DOCUMENT RESUME

ED 187 806 UD 020 610

AUTHOR	Goodrich, Robert L.; St. Pierre, Robert G.
TITLE	Opportunities for Studying Later Effects of Follow Through.
INSTITUTION	Abt Associates, Inc. Cambridge, Mass.
SPONS AGENCY	Department of Health, Education, and Welfare, Washington, D.C.
REPORT NO	AAI-R-79-156
PUB DATE	Dec 79
CONTRACT	HEW-300-78-0443
NOTE	141p.; For related documents see UD 020 611-614.
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Academic Achievement; Disadvantaged Youth; *Educational Research; *Longitudinal Studies; Outcomes of Education; Primary Education; *Program Effectiveness; *Research Needs; Student Participation
IDENTIFIERS	*Follow Through

ABSTRACT

This study assessed the feasibility of conducting relatively low cost later effects studies of Follow Through (FT) programs without extensive new data collection, relying instead on existing data resources including the national FT data base, sponsor and site data bases, school and community records, and potential new data collection of modest scope. Six data sets were selected as meeting the criteria (adequate sample size, availability of a wide range of measures, and good local comparison groups). The following research questions were identified as being suitable for study using existing data bases: (1) What are the causal relationships among academic achievement, IQ, process measures, and affective measures? (2) Do achievement effects associated with FT participation increase, decrease, or stabilize after program completion? (3) What is the relative importance of the current academic year and the child's previous academic history as determinants of academic success? (4) What are the causal relationships between academic performance and school attendance in and after FT? (5) What is the relationship between post-FT academic performance and length of FT participation? (6) Is it possible to predict post-FT patterns on different measuring instruments? (7) What differences in post-FT performance exist among different models at the same site? (Author/MK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED187806

ABT ASSOCIATES INC.
55 WHEELER STREET, CAMBRIDGE, MASSACHUSETTS 02138
TELEPHONE • AREA 617-492-7100

AAI Report No. 79-156

Opportunities for Studying Later Effects of Follow Through

by

Robert L. Goodrich
Robert G. St.Pierre

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

December, 1979

This report is made pursuant to Contract HEW-300-78-0443. The amount charged to the Department of Health, Education, and Welfare for the work resulting in this report (inclusive of the amounts so charged for prior Later Effects Study reports submitted under this contract) is $163,422. The names of the persons, employed or retained by the contractor, with managerial or professional responsibility for such work, or for the content of the report are as follows: Robert G. St.Pierre and Robert L. Goodrich.

UD020610

Robert G. St.Pierre Contract Manager	David C. Hoaglin Quality Control Reviewer	Management Reviewer

MAY 9 1980

STAFF OF AN EXAMINATION OF ALTERNATIVES FOR FOLLOW THROUGH EXPERIMENTATION

Abt Associates Inc. Staff

Project Management

Robert G. St.Pierre, Project Director
Elizabeth C. Proper, Director of the Implementability Study
Robert L. Goodrich, Director of the Later Effects Study

Field Staff

Nancy L. Ames
Richard B. Anderson
D. Catherine Baltzell
David Boesel
Nancy Brigham
John Doucette
A. Tara Feraco
Nancy N. Goodrich
Michael H. Hennessy
Karen Keefe
Yevgenia D. Mackiernan
Bernadette Nelson
Eugene C. Royster
Linda B. Stebbins
D. Susan Thompson

Support Staff

Yevgenia D. Mackiernan
Sandra Margolin
Mary Ellen Perry

In-House Advisors

David Hoaglin
Karen S. Louis
Sheila Rosenblum

Abt Associates Consultants

Implementability Study

Richard Elmore
Michael Fullan
Alan Pomfret

Later Effects Study

Jacob Cohen
Thomas D. Cook
Walter Haney
Leslie McLean

U. S. Office of Education Staff

Project Officers

Burma H. Hulten
Daniel Ozenne
Gerald Burns

Table of Contents

Page

1.0 INTRODUCTION . . . 1
1.1 Summary of Recommendations and Research Questions. . . 2

2.0 BACKGROUND . . . 7
2.1 History of FT Later Effects Studies . . . 7
2.2 Rationale for FT Later Effects Studies . . . 9

3.0 EXISTING STUDIES OF LATER EFFECTS . . . 11
3.1 Arizona (Wichita, KS). . . 11
3.2 Arizona (Lakewood, NJ) . . . 12
3.3 Bank Street (New Haven, CT). . . 13
3.4 Far West (St. Louis, MO) . . . 15
3.5 Kansas (Sponsor-wide). . . 15
3.6 Oregon (East St. Louis, MO). . . 16
3.7 Oregon (Sponsor-wide). . . 17
3.8 Summary of Existing Studies. . . 17

4.0 IDENTIFICATION OF DATA BASES SUITABLE FOR LATER EFFECTS STUDIES. . . 19
4.1 Usefulness of the National Evaluation Data Base for Studying Later Effects. . . 19
4.1.1 Use of the National Evaluation Data Base as a Roster File. . . 19
4.1.2 Use of the National Evaluation Data Base to Supplement Sponsor-Supported Data Sets. . . 20
4.2 Census of Sponsor-Maintained Data Sets . . . 21
4.3 Criteria for Selection of Sponsor/Site Data Sets . . . 26
4.3.1 The Level of Analysis. . . 26
4.3.2 Bases of Comparison. . . 28
4.3.3 Attrition. . . 30
4.3.4 Dependent Measures . . . 31
4.3.5 Statistical Power. . . 33
4.3.6 Data Quality . . . 34
4.3.7 Measures of Post-FT Curriculum and Process . . . 35
4.4 Qualifying Data Sets . . . 36

5.0 DISCUSSION OF DATA BASES SUITABLE FOR LATER EFFECTS STUDIES. . . 41
5.1 General Analytic Approach. . . 41
5.1.1 Preliminary Exploratory Analyses . . . 41
5.1.2 Summative Analyses . . . 47
5.1.3 Validation of Conclusions. . . 49
5.2 Fordham University Data Base . . . 51
5.2.1 Educational Model and Later Effects Hypotheses . . . 51
5.2.2 General Description of the Data Base . . . 51
5.2.3 Preliminary Study Plan . . . 60
5.3 University of Oregon Data Base . . . 66
5.3.1 Educational Model and Later Effects Hypotheses . . . 66
5.3.2 General Description of the Data Base . . . 66

Table of Contents
(continued)

Page

5.3.3 Preliminary Study Plan 74
5.4 University of Arizona (Wichita, KS) Data Base. 82
5.4.1 Educational Model and Later Effects Hypotheses 82
5.4.2 General Description of the Data Base 82
5.4.3 Preliminary Study Plan 86
5.5 Southwest Educational Development Laboratory Data Base 88
5.5.1 Educational Model and Later Effects Hypotheses 88
5.5.2 General Description of the Data Base 88
5.5.3 Preliminary Study Plan 91
5.6 Georgia State University Data Base 96
5.6.1 Educational Model and Later Effects Hypotheses 96
5.6.2 General Description of the Data Base 96
5.6.3 Preliminary Study Plan 98
5.7 School District of Philadelphia Data Base. 100
5.7.1 Educational Models and Later Effects Hypotheses. 100
5.7.2 General Description of the Data Base 102
5.7.3 Preliminary Study Plan 109
5.8 Integration of Studies 115

References 117

Appendix 1: The Linear Discriminant Function as a Composite Covariate . 123

List of Tables

Page

Table 1 Grade in 1979-80 by National Evaluation Cohort 20

Table 2 Sponsors with Insufficient Data. 22

Table 3 Sponsors with Site-Level Data 23

Table 4 Sponsors with Multi-Site Data. 24

Table 5 Sites with Multi-Sponsor Data. 25

Table 6 Summary Assessment of Data Bases 37

Table 7 Sample Size by Measure, Cohort, Grade, and Test Point for the Fordham (New York) Data Base 53

Table 8 Sample Size by Measure, Cohort, Grade, and Treatment for the Fordham (Atlanta) Data Base. 57

Table 9 Measures Available on the Fordham Data Base by Grade and Measure (Atlanta Project, Cohort IV) 62

Table 10 Estimated Resources Required for the Proposed Fordham Study. 67

Table 11 Measures Available on the Oregon Data Base by Cohort and Grade. 69

Table 12 Measures Available on the Oregon Data Base 70

Table 13 Relation of Oregon Objectives to Measures on Oregon Data Base. 71

Table 14 Comparison of FT and NFT Groups on Selected Fifth Grade Demographic Variables by Site for the Oregon Data Base. 73

Table 15 Longitudinal Sample Size by Site and Grade for the Oregon Data Base 75

Table 16 Estimated Resource Requirements for the Proposed Oregon Study 81

Table 17 Test Data on the Wichita Data Base 83

Table 18 Sample Size by Grade, Cohort and Treatment Group for the Wichita Data Base. 85

List of Tables
(continued)

Page

Table 19 Estimated Resource Requirements for the Proposed Wichita Study 89

Table 20 Sample Size at Entry to FT and Ethnic Composition by Site for the SEDL Data Base 91

Table 21 Estimated Resource Requirements for the Proposed SEDL Study 95

Table 22 Sample Size by Grade, Treatment, and Cohort for the Georgia State Data Base 97

Table 23 Estimated Resources Required for the Proposed Georgia State Study 101

Table 24 Test Data on Merged National Evaluation/Philadelphia File 103

Table 25 Process Data by Grade Available on the Philadelphia Data Base 104

Table 26 Background Data by Cohort, Stream and Grade for the National Evaluation Data Base--To Be Merged with the Philadelphia Data Set 105

Table 27 Estimated Resouces Required for the Proposed Philadelphia Study 116

List of Figures

Page

Figure 1 Summary of Research Questions that Can Be Addressed by Each Recommended Data Base 5

1.0 INTRODUCTION*

This document is a feasibility and planning report for studies of the effects of Follow Through (FT) models on children in the post-FT grades. Its purposes are to discuss the rationale for "later effects" (LE) studies, to document the activities performed in Abt Associates' (AAI) feasibility and planning study, to review the characteristics, content, and usefulness of existing data bases for post-FT analyses, and to identify data sets suitable for LE studies.

This is the last in a series of three reports. The first (Goodrich and St.Pierre, March 1979) examined the utility of the FT national evaluation data base for supporting LE studies. The second (Goodrich and St.Pierre, June 1979) presented a summary of existing data bases maintained by FT sponsors and sites. Both reports concentrated on fairly mechanical issues of data availability and technical usefulness rather than on detailed articulation of research questions, the importance and usefulness of expected results, cost estimates, or specific analytic strategies. These latter topics are covered in the current report. In addition to the introduction, the following sections are included in this report:

- Background. We discuss the history of LE studies in FT, the rationale for LE studies, and the potential for building LE studies into future FT research plans.
- Existing Studies of Later Effects. We review a selection of existing LE studies conducted by FT sponsor and site personnel in order to establish expectations for findings of new studies and to identify areas where more work is needed.
- Identification of Data Bases. We define the purpose of and the constraints on the present feasibility study, review the methodology of the present study, discuss criteria for selection of data sets, and identify a set of data bases that can support LE studies.
- Discussion of Recommended Studies. We present detailed characteristics of data bases that can support LE studies and offer preliminary analytic plans.

---

*The research reported herein was performed pursuant to Contract No. HEW-300-78-0443 with the U.S. Office of Education. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official U.S. Office of Education position or policy.

1.1 Summary of Recommendations and Research Questions

Although post-FT data were not collected as part of the national evaluation, several FT sponsors and sites collected such data in their own evaluation efforts. The present study was designed to assess the feasibility of conducting relatively low-cost post-FT studies without extensive new data collection, relying instead on existing data resources, including the national evaluation data base, sponsor-maintained data bases, site-maintained data bases, school and community records, and potential new data collection of modest scope.

The constraint of relying on accessible, existing data limits the range of analytic approaches that can be suggested. We cannot recommend any single design or approach, but must use data bases as they exist and attempt to build the best possible studies. The range of studies that we have considered has also been limited by cost considerations. For example, almost all FT sites have data that could support some form of LE study. If sufficient time and funding were available, it would be possible to access school district files and construct data sets of varying quality from most sites. We have not explored this possibility; instead, we have concentrated our efforts on sponsors and sites where the data are accessible, are of higher quality, and require minimal additional preparation for analysis.

At the beginning of this project, we decided that the best way to identify high-quality data bases was to follow an iterative process in which we first selected a large number of data bases that could potentially support LE studies, learned enough about those data bases to select a subset for further work, obtained more information, selected a smaller set to investigate even further, and so on. We initially conducted a census of sponsors and inquired as to the existence of data at the sponsor and site levels. From these, we selected eleven data bases that might support LE work. After further investigation, we have identified six data bases that meet our criteria for LE studies. Our opinion is that the national evaluation data base is not useful by itself for post-FT studies, but that it can be used as an additional data resource for appending to sponsor and site data bases, and as a roster file for locating children.

Two of the sponsor/site data bases afford especially good opportunities, because they surpass considerably the minimum methodological standards (i.e., adequate sample size, availability of a wide range of measures, good local comparison groups), and because findings might be generalizable to populations of interest. Along with a multi-sponsor study made possible by using the data set maintained by the Philadelphia Public Schools, we have a total of three top-ranked data bases. These are as follows:

1. Philadelphia, merged with the national evaluation data base, concentrating on Cohorts I-III;
2. Oregon, concentrating on five sites with well-matched local control groups; and
3. Fordham, including both Atlanta, GA and New York City.

Studies using these three data bases would almost certainly develop new and useful knowledge about FT and should be pursued if at all feasible.

The remaining data bases that we recommend for study are weaker methodologically and/or in generalizability. They focus on highly specialized single sites with marginal sample sizes. These are:

4. Arizona (Wichita, KS);
5. Georgia State (Natchitoches Parish, LA); and
6. Southwest Educational Development Laboratory (San Diego, TX; possibly Los Angeles, CA and St. Martin Parish, LA).

Any of the latter three data bases will support smaller studies that, since they involve only one site each (SEDL possibly excepted), would not allow general conclusions to be drawn about the educational models involved. They would be useful mainly as important quantitative case studies. Together with the other studies, however, and with sponsor/site studies, they may help substantially in fitting together a more complete picture of FT.

In planning new LE studies, one would typically begin with the important research questions and then develop a design that ensures availability of the necessary data. In the present case, the data already exist, and the question is whether they can support useful LE studies. Our general approach has been to concentrate on the availability and quality of the data first, rather than on the research questions they might address.* This approach can be thought of as a technical survey of the research opportunities for new knowledge about post-FT effects. The earlier reports by Goodrich and St.Pierre (March 1979, June 1979) achieve this goal by discussing characteristics of several candidate data bases. In the present report, we go on to indicate which of several research questions can be addressed by each of the six recommended data bases. Figure 1 summarizes the major questions that can be addressed by post-FT studies based on existing data, and shows which questions can be addressed through which data sets. It is clear that the first three data sets, supported by Philadelphia, Fordham, and Oregon respectively, offer the chance to address a wider range of questions than would be possible with the other data bases. Still, the latter three data sets (Arizona, Georgia State, SEDL) can provide valuable information on a limited number of research questions.

---

*One of our consultants, Dr. Walter Haney, believes that this approach is backward and that we should have begun with the research questions. We feel, though, that our approach is appropriate because it concentrates on the scope of opportunities rather than on the usefulness of the data to provide answers to what may be a narrow range of predetermined questions.

As part of an overall analysis strategy, analyses of all data sets that meet the criteria we set could be conducted independently. The results of this work would generalize only to the specific data subsets. The findings could then be integrated to the sponsor level (treating them essentially as data points for meta-analysis), and to the national FT level to the extent possible. The final product, a series of LE studies, could then be an integrated picture pieced together from available evidence. We suggest that the integrative analysis include results from existing sponsor and site analyses that meet acceptable standards of methodological rigor.

Since we have considered only studies based principally on existing data, they will be considerably less expensive than post-FT studies involving substantial new data collection. They can also provide information to USOE within one or two years, whereas studies involving new data collection from FT children beyond the third grade could extend for as long as a decade. By avoiding a multi-year period of data collection, evaluators can provide USOE with timely information and can provide other researchers with an important base of information for the design of entirely new studies.

Figure 1

Summary of Research Questions that Can Be Addressed by Each Recommended Data Base

RESEARCH QUESTIONS	RECOMMENDED DATA BASES					
	Philadelphia	Fordham	Oregon	Arizona	Georgia State	SEDL
1. What are the causal relationships among academic achievement, IQ, process measures (amount of FT exposure, attendance, etc.), and affective measures?	X	X	X			
2. Do achievement effects associated with participation in FT increase, decrease, or stabilize after completion of the program?						
a. If positive effects on academic achievement are produced by FT models at the end of grade 3, do they fade out, or are they sustained in the post-FT years?	X	X	X		X	
b. Do FT models that aim for achievement gains by first changing children's affect, conceptual ability, parental behavior, and so on, produce "sleeper effects" on achievement measures?	X					
c. How does post-FT academic performance relate to previous FT performance and initial status (e.g., socioeconomic status, pretest, demographic data)? Are gains sustained in the post-FT years?	X	X		X	X	X
3. What are the relative importances of the current academic year and the child's previous academic history as determinants of academic success?	X	X	X			
4. What is the causal relationship between academic performance and school attendance? In FT? After FT?	X	X	X			
5. What is the relationship between post-FT academic performance measures and length of FT exposure?	X	X		X		
6. Is it possible to predict post-FT patterns (fade-out, sleeper effects, sustained growth, etc.) on different measures (e.g., math, reading, Metropolitan Achievement Test, Wide Range Achievement Test) in terms of the FT curriculum?	X		X			
7. What differences in post-FT performance (achievement, grades, attendance) exist among different models at the same site?	X					

2.0 BACKGROUND

The purpose of this section is to discuss the history of LE studies in FT, the rationale for studies of post-FT effects, and the future of LE studies in FT.

2.1 History of FT Later Effects Studies

The long-term effects of educational programs is an important but neglected area of study. In order to provide information to policy makers within a reasonable time frame, evaluators of educational programs have traditionally measured impact at the end (or even before the end) of the treatment. Studies of post-treatment effects have been seen as interesting but not timely, since, as Haney (1977) put it, "...the policy maker can hardly be expected to defer evaluation to his or her grandchildren" (p. 178). However, the original purpose of FT was to improve life chances of low-income children, and surprisingly little attention has been paid to post-FT effects, especially considering that concern over the apparent fade-out of Head Start effects (Wolff and Stein 1966) was the primary impetus behind the establishment of FT.

The amount of attention given to the study of post-FT effects has differed across three levels of evaluation: national studies funded by USOE and conducted by contractors; studies carried out by the FT sponsors; and studies conducted by evaluation personnel in FT sites. Although no investigation of post-FT effects has been funded at the national level, proposals for the study of such effects are not new. One proposal by Madaus and Marx (1973) of the Huron Institute detailed plans for a "later grades study" structured so that it would follow children from the national evaluation. Most of the research issues cited by Madaus and Marx are still valid: whether achievement gains associated with FT participation increase, stabilize, or decrease over time; whether achievement gains associated with various FT models are differentially affected during the post-FT years; the possible ways in which teachers, administrators, and parents might be affected by the transfer of FT children into standard fourth grade classrooms; and the extent of diffusion of FT practices into non-Follow Through (NFT) classrooms. They were also in favor of maintaining the data base at the end of the national FT evaluation, as it can serve as a valuable resource for future educational research. The work suggested by Madaus and Marx in 1973 was not funded, and the opportunity to gather continuous post-treatment data on children completing the FT program in 1974 and 1975 was lost.

A suggestion that post-FT effects be studied at the national level was again made in 1976, when USOE requested the Huron Institute to identify a set of FT evaluation research studies that might be pursued on a short-term basis. One of the twenty-five preliminary study suggestions made in a draft report by Garet (1977) was a plan to study the effects of FT in the later grades. The general purpose of this plan was to explore hypotheses about persistence or fade-out of third grade effects in the later grades and delayed effects that might appear subsequent to completion of the FT treatment, particularly on nonstandardized measures. Garet recommended that a contractor develop a battery of instruments focusing on academic achievement, student interviews, and productive language, perhaps including portions of the National Assessment of Educational Progress battery. The battery would be administered to fourth, fifth, and sixth grade children who did and did not participate in FT, and the focus would be on whether post-FT results supported or challenged with the national evaluation's end-of-third-grade findings. In Garet's (1977) final report, the twenty-five preliminary research suggestions were reduced to ten and no longer included the later grades study.

At the sponsor level, interest in post-FT effects has existed for several years. Debates about persistence of effects, fade-out, and sleeper effects have taken place regularly. Claims of delayed effects on achievement have often been made by proponents of the more open FT models as partial justification for the absence of effectiveness in the early grades. Some sponsors (e.g., Becker and Engelmann 1978; Ramp, Williams and Branden 1978) have collected data and conducted their own studies of the achievement test performance of graduates from their respective programs.

Interest in LE studies is also evident at the local level. Empirical studies have been conducted by local evaluators, sponsor personnel, doctoral students, and others. In spite of its potential importance, this work has received relatively little attention to date--perhaps because of its relative inaccessibility. A brief review of several sponsor- and site-level research articles is included in Section 3 of this report. These articles are the principal resource for current knowledge of post-FT effects.

2.2 Rationale for FT Later Effects Studies

With this history in mind, we see the following as major reasons for studying post-FT effects. First, FT was initiated because the achievement gains obtained by Head Start children appeared to dissipate in the early school years. The same issue should be investigated with respect to FT: if achievement gains are produced by FT, do they dissipate, or are they sustained after completion of the program?

Second, many of the FT sponsors try to influence childrens' affect, conceptual ability, cognitive style, and other non-achievement areas, as well as parental behavior. These sponsors believe that achievement gains will be obtained indirectly through impacts on mediating variables such as those listed above. In this case, effects on student achievement are not expected to occur quickly. Therefore, in order to evaluate these sponsors' models, it is necessary to examine achievement gains subsequent to program termination. The question is, do FT models that aim for achievement gains by first influencing these mediating variables produce delayed or "sleeper" effects?

A third motivation for studies of post-FT effects is that FT projects, by design or as a byproduct, may affect parental and community opinions. In general, parents support FT and have had considerable impact on the decision to maintain the program in the face of efforts to phase it out. The question that might now arise is, what are parents', teachers', and students' opinions of FT once students have made the transition to "regular" public school classes?

Fourth, FT's overall objective has been to improve the "life chances" of low-income children. College attendance, future earnings, social mobility, and other post-school variables are all important indicators of success in life but are not available measures in any LE study now contemplated. Life chances studies would concentrate on earlier, accessible indicators of success such as school attendance, special education placement, grade retention, dropout rate, grades, course selection, and discipline records. Our discussions with consultants and FT stakeholders indicate great interest in FT effects within this domain of post-FT variables.

LE work can play an important part in the future of FT. First, such empirical studies can address certain policy questions. For example, should FT be extended to later grades (as Head Start was extended via FT)? If so, should all models be extended, or only those that are found to be successful in maintaining positive end-of-third-grade effects or having positive sleeper effects?

Second, the knowledge gained through the present project and through any actual LE studies funded by USOE will be of value in planning future FT research. To the extent that current LE work suggests that FT models are or are not having positive post-third-grade effects, future studies can be designed to capture similar kinds of information more completely and accurately.

The nature of our work is therefore exploratory. We think it appropriate to regard the results from any LE studies based on extant data as being tentative and suggesting directions and hypotheses for future work, rather than as being definitive or summative. Although an attempt should be made to integrate the results of any new studies with existing analyses across all sponsors, the ability to generalize will be quite constrained, and the results of the recommended work are likely to be more appropriate for generating than for confirming hypotheses.

Finally, future FT research should consider studying post-FT effects in a context larger than that of the variables discussed in this paper (child achievement test scores, attendance, dropout rate, parent/student interviews, etc.). For example, the degree to which FT sponsors are able to institutionalize their models so that children, parents, and communities benefit even if FT funds and sponsor assistance are no longer available is a prime area of LE study that has not yet been considered. Other topics not commonly studied also fall into the "delayed effect" category. For example, diffusion of FT benefits to siblings of FT children, diffusion of FT benefits through the exposure of non-FT children to FT-trained teachers, increases in teacher competencies through FT training, increases in the life chances of FT parents through FT-supported employment, career development and other forms of parental involvement, and general curriculum development are all of interest as potential study topics.

3.0 EXISTING STUDIES OF LATER EFFECTS

A number of reports by FT sponsors and sites should be considered in defining an overall strategy for studying post-FT effects. Several contribute substantially to the growing body of research on FT. The purpose of this section is to present synopses of the reports we have read and our judgments concerning their methodological soundness. We do not claim that all LE reports of value are included here. However, a review of recently completed efforts will help to establish expectations for the range of results that might emerge from future LE studies, will help generate preliminary hypotheses for new studies, and will also illuminate the areas where more work is needed. The review includes synopses of studies concerning Arizona (Wichita, KS), Arizona (Lakewood, NJ), Bank Street (New Haven, CT), Far West (St. Louis, MO), Kansas (sponsor-wide), Oregon (East St. Louis, MO), and Oregon (sponsor-wide).

3.1 Arizona (Wichita, KS)

Riley, G.R. "Follow Through Program Evaluation, 1976-77." Wichita P.S. Division of Research, Planning and Development Services, January 1978.

In a report prepared for the Wichita Public Schools, Riley (1978) compiled data collected as part of a local evaluation of the Tucson Early Education FT Model sponsored by the University of Arizona. While the Wichita data set is relatively small, it extends across several cohorts and possesses many of the features of classical experimental designs; e.g., treatment and control groups were selected randomly from a common pool of eligible children and were followed from kindergarten through grade 6. About 100 FT children and 100 NFT children were selected each year.

Several groups of children were followed in this study. A list of Head Start-eligible children was drawn up, based on social worker assessments of educational need. The "Head Start/FT" group was selected from a randomly ordered list of Head Start-eligible children, working down from the top of the list until 100 children were selected. The "Head Start/NFT" group consists of all children remaining on the list, and includes children who rejected FT (only a minor departure from full randomization when rejection rates are low). The Head Start/NFT group had a larger proportion of whites until 1977-78, when the proportion of FT refusers jumped substantially. This occurred because a policy was implemented whereby all FT children were to be bused out of their neighborhoods. The "FT classmates" group consists of non-eligible classmates of Head Start/FT children who received the FT classroom treatment but not the comprehensive services. The "low-income" group were on the waiting list for Head Start and are less needy than Head Start children. The "non-low-income" group are randomly selected non-low income children in four paired non-low income schools.

Data have been collected via several instruments, the most important of which are the Illinois Test of Psycholinguistic Abilities, administered as a pretest at entry to kindergarten, and the Iowa Tests of Basic Skills, administered in the spring in grades 3-6. Standard demographic data such as sex, age, and race are also in the data base. At completion of the 1978-79 school year, five cohorts will have completed the K-3 FT program and will have been followed longitudinally through grade 6 and beyond. Longitudinal sample sizes average about thirty-five Head Start/FT and thirty Head Start/ NFT children per cohort. Roughly similar sample sizes exist for the other groups.

Summary of Findings

The 142-page evaluation report is almost totally composed of tables, figures and other data displays. The main form of analysis is the t-test, which has been used to compare the five treatment groups on various outcomes by cohort and year. The results are undigested. There is little discussion, and readers are left to draw their own conclusions about the effectiveness of FT beyond the lowest level of aggregation (e.g., across measures, cohorts, or years). In addition, very little is said about the relationship of post-FT effects to earlier effects. Thus, the report by Riley (1978) is, in itself, a source of data for further study.

From our reading of the reports the results are discouraging. We see few significant effects and, of the few, more than one-half favor NFT. Overall, there is little evidence of educationally significant FT effects at grade 3 or later for participants in the Tucson Early Education FT model. Since FT effects are small, fade-out is not an issue. "Sleeper effects" do not appear to exist in this study.

3.2 Arizona (Lakewood, NJ)*

Cloud, K., R.K. Rentfrow, and J. Hildebrandt. "Later School Success of Follow Through Graduates in Lakewood New Jersey Public Schools. Tucson, AZ: Center for Educational Research and Development, July 1979.

---

*Much of the study description is taken from an abstract written by Cloud et al. (1979).

Cloud et al. investigated the later school success of the first group of low-income pupils enrolled in the FT program in Lakewood (NJ) public schools. These children participated in both Head Start and FT, and thus had five years of integrated early childhood education. A comparison group was formed from older siblings of these FT pupils, and the school records of these two groups in grade 10 were compared. Data were available for twenty-five pairs of children. The variables investigated were: school attendance at grades 4, 7, and 9; course grades in English, math, and science at grade 9; retention; special education placement; drop-out rate; and educational aspiration.

Summary of Findings

Results indicated that these two groups differed on two variables. The FT graduates had a significantly lower rate of grade retention than their sibling comparison group. Eleven of the twenty-five siblings had been retained compared with no retentions for the FT group. The FT group also had a lower incidence of special education placement than their sibling comparison group. Five of the twenty-five siblings had been placed in special education, compared with two in the FT group. A similar result was found in drop-out rate, with the FT group having a lower rate, but the difference was not significant.

The Cloud et al. study is especially interesting in that the findings replicate the work of the Developmental Continuity Consortsium (Lazar et al., 1977). The study argues for including "life chances" variables in LE studies if at all possible.

3.3 Bank Street (New Haven, CT)

Seitz, V., N.H. Apfel, and C. Efron. "Long-Term Effects of Early Intervention: The New Haven Project," paper presented at 1977 annual meeting of American Association for the Advancement of Science, Denver, Colorado.

Seitz, Apfel and Efron (1977) studied post-FT achievement and IQ test scores by tracing children from the Abelson (1974) study of FT effects. The Seitz study includes two cohorts beginning FT in 1967 and 1968 as part of the New Haven Bank Street project. Longitudinal samples include about thirty to thirty-five FT and thirty to thirty-five NFT low-income children in each cohort. The Cohort I comparison group was selected in kindergarten, but the Cohort II comparison children were selected in grade 3.

Data collected at entry to the study include the Peabody Picture Vocabulary Test and the Screening Test of Academic Readiness. Standard background data were also collected; e.g., child's age, mother's education, number of siblings. FT and NFT children from both cohorts were tested at the end of grade 3 with the Peabody Picture Vocabulary Test and the Peabody Individual Achievement Test. The same tests were used in grades 4-8, but the time of administration varied by cohort.

Summary of Findings

Seitz et al. found math, general information, and Peabody Picture Vocabulary Test IQ effects favoring FT for boys in Cohort I and math, spelling, general information, Peabody Individual Achievement Test, and Peabody Picture Vocabulary Test IQ effects favoring FT for girls in Cohort II. No other significant effects were found. Hypotheses were tested using one-tailed t-tests. The authors claim that the most interesting findings in the study were "...(a) the apparent absence of a 'fade-out' effect for those groups of intervention children who were significantly ahead of non-intervention children at the end of the program, and (b) the evidence for increasing differences, or a sleeper effect, for girls after the intervention had ceased." (Seitz et al., p. 21).

This study is seriously flawed, and the results are questionable. Sample sizes are low, and the longitudinal NFT sample does not seem to be equivalent to that for FT. Comparisons of the FT and NFT Cohort I longitudinal children at entering-K show that mothers of FT boys were significantly better educated than mothers of NFT boys, and that the FT boys scored significantly higher in IQ (90.8 vs. 69.7) than NFT boys. The Cohort I girls were better matched. Entering-K measures are not available for Cohort II NFT children, but entering-K comparisons of FT children by sex are interesting: in Cohort I, boys had significantly higher IQ's than girls, and in Cohort II, girls had superior IQ's.

Seitz's effects were found just where one would expect on the basis of entering-K IQ scores; most of the results are likely consequences of prior demographic differences. With samples as small as those in the New Haven study, we must expect bad matches between FT and NFT on the basis of chance alone, and even though these differences are usually not statistically significant (because of the lack of statistical power), they are practically significant. Failure to reject the null hypothesis has almost no meaning in the New Haven study.

We do not believe that there are any credible findings in the Seitz study.

3.4 Far West (St. Louis, MO)

Bridewell, J.L. and R.H. Edwards. "Follow Through 1977-1978," St. Louis Public Schools, 1978. "Following Through on Follow Through," paper presented at 1979 meeting of the American Educational Research Association, San Francisco, April 1979.

Bridewell and Edwards (1979) provide a particularly interesting LE study. Their motivation for studying post-FT effects was that St. Louis was identified in the national evaluation (Stebbins et al. 1977) as a site in which the Far West Model performed particularly well. Yet, local evaluators had never found the program to be especially successful. This apparent contradiction spurred the researchers to consider potential sources of the national/local conflict. In doing so, they constructed a data base on FT and NFT children in Cohorts I-IV containing standardized test results on reading, language, and arithmetic through grade 5. Longitudinal sample sizes were about 125 FT and 2400 NFT children, and the bases of comparison are the district norm and the publisher's standard (grade equivalent).

The major weaknesses of the study are that results depend on the grade equivalant metric and on the assumption that district averages are a valid base of comparison. The authors are also worried about possible test administration/scoring irregularities. The study is currently being extended to include sixth grade data.

Summary of Findings

There is evidence for the acceleration of achievement in all three areas, beginning in grade 4. The report suggests the possibility of a delayed or "sleeper" effect, although the authors do not claim to have proved anything.

3.5 Kansas (Sponsor-wide)

Assorted reports and materials prepared by the Kansas FT office.

Kansas has prepared a number of cohort-by-cohort mean achievement growth charts for reading and arithmetic scores on the Wide Range Achievement

Test. The basis of comparison is usually the publisher's test norm, and the metric is the grade equivalent.

Summary of Findings

These charts appear to show consistently that reading scores continue to grow in fourth and fifth grades at normal or close to normal rates, but that arithmetic scores decelerate sharply in post-FT kneoh. Positive effects in reading are evident in fourth and fifth grades. A few comparisons with NFT groups are consistent with the results based on norms. Overall, the Kansas data support a fade-out interpretation; i.e., the program secures end-of-third grade effects, which persist for a time but eventually fade out.

3.6 Oregon (East St. Louis, MO)

Weber, B. and M. Fuhrmann. "A Study of District A's Former Follow Through Students' Retention of Basic Skills After Six Years Out of the Program," April 1978. Paper prepared for "District A."

Weber and Fuhrmann (1978) conducted a study of the basic skills achievement of students who, from 1968-69 through 1970-71 participated in the grades 1-3 FT program sponsored by the University of Oregon in East St. Louis, Missouri. Two hundred forty students entered the FT program in grade 1, 187 of these completed FT in grade 3, and 120 were traced through grade 9. A post hoc comparison group of 120 ninth graders who had not been in FT was randomly selected from the district-wide pool of approximately 1400 students. Data for the study came from the district-wide testing program, in which all ninth graders were tested with the California Achievement Tests.

Summary of Findings

Analysis of grade equivalent scores at grade 9 showed that FT graduates scored significantly higher than comparison children on all subtests of the California Achievement Tests. FT/NFT differences ranged from five months on the total test score to eight months on the reading subtest. When converted to raw scores and reanalyzed, the data showed significant FT-favoring differences on all subtests except spelling.

The major flaw in the study is that a posttest-only design was used to assess the hypothesis that, six years after completion of the program, FT students would "retain the basic academic skills at a higher level" than NFT students. To "retain" skills implies that some measure of FT effect was taken at completion of the FT program in grade 3. This was not the case, so it is not clear whether the FT/NFT effects observed in grade 9 represent increases, decreases, or no difference from the unmeasured grade 3 FT/NFT effect.

Because of the faults cited, we would not draw conclusions from this report.

3.7 Oregon (Sponsor-wide)

Becker, W.C. and S. Engelmann. "Analysis of Achievement Data on Six Cohorts of Low-Income Children from Twenty School Districts in the University of Oregon Direct Instruction Follow Through Model," Technical Report 78-1, University of Oregon Follow Through Project, December 1978.

Becker, W.C. and R.M. Gersten, Appendix 2 to Report 78-1, April 1979.

These Oregon researchers analyzed fifth and sixth grade Wide Range Achievement Test (Level 1, Level 2) and Metropolitan Achievement Test reading, math, spelling, and science scores in 1975 and 1976. Six (1976) and seven (1975) sites with good NFT comparison groups were included. They present charts depicting raw scores with hypothesis tests based on analysis-of-covariance-adjusted scores. The measure is the normal curve equivalent score, but charts are labeled by percentile. Analyses are site by site.

Summary of Fingings

Overall, many persistent positive effects in fifth and sixth grades are observed on both the Wide Range Achievement Test reading score and the Metropolitan Achievement Test reading subtest. There are scattered positive effects on the Metropolitan Achievement Test science and spelling subtests. Few persistent positive math effects are present. There are not many signs that sixth grade performance is diminished from fifth grade performance.

3.8 Summary of Existing Studies

The four "motivations" for LE studies cited in Section 2.2 have to do with finding evidence of persistence of academic effects, sleeper effects, persistence of attitudinal effects, and life chances. Sponsor and site research has concentrated on persistence of academic effects (fade out), and sleeper effects. One study investigated the effects of FT on life chances variables. Our reviews of these studies, together with the results of many conversations with FT sponsors, support the following conclusions.

First, models that produce FT effects by grade 3 appear to exhibit fade-out, particularly in math. This conclusion is based principally on research carried out by the Oregon and Kansas sponsors, and is a relatively strong one.

Second, there is no strong evidence for sleeper effects, although some work points in this direction, particularly in the Far West data and the New Haven (Bank Street) data. These "tidbits" only whet the appetite for further study.

Third, a study of life chances variables showed positive effects on retention and special education placement. The fact that this study replicates the findings of the Developmental Continuity Consortsium lends credence to the argument that the long-term effects of FT should be measured in terms of life chances variables as well as in terms of academic achievement.

It might be asked whether, given the meagerness of conclusions that can be drawn from existing studies, new studies based on old data are likely to yield much information. In fact, we think that possibilities are very good. Nearly every sponsor interviewed cited limited resources as the major reason for the lack of analysis of existing data. That is, there is a large amount of LE data available that simply have not yet been analyzed and that can bear directly on the research questions cited in Section 1.1.

4.0 IDENTIFICATION OF DATA BASES SUITABLE FOR LATER EFFECTS STUDIES

In this section, we discuss the procedure used to identify sponsor/site data bases as resources for potential new studies of post-FT effects.

4.1 Usefulness of the National Evaluation Data Base for Studying Later Effects

One potential resource for studying post-FT effects is the data base resulting from a ten-year, large-scale, USOE-funded national evaluation of FT. We began our review of the national evaluation data base (maintained by SRI International) by assessing the knowledge of the structure and content of the data that we gained through our experience with the national evaluation. We extended this knowledge by reviewing SRI's most recent data base documentation (Middleton and Durgin 1978), and by holding telephone conversations with Dr. Teresa Middleton of SRI. Then we considered the specific ways that the national evaluation data base could be used to overcome or reduce defects of sponsor and site data, and/or to stand alone as a data resource for LE studies.

Because the national evaluation data base does not contain post-FT measures, it is not in itself useful for LE studies. Rather, it provides a resource file of background variables, achievement measures, and a limited range of non-cognitive measures collected during the FT school years. The data can be used in two ways. First, the national evaluation data set can be regarded as a roster to help locate former FT children, teachers, administrators, or parents for studies based on measures that exist outside the national evaluation data base. Second, the national evaluation data can serve as a source of cognitive and non-cognitive measures for grades K-3, to be used as statistical controls or as independent variables in LE studies based on sponsor- or site-collected data. The next two sections discuss these two potential uses in more detail.

4.1.1 Use of the National Evaluation Data Base as a Roster File

The national evaluation data base can indeed serve as a resource to identify samples of students, parents, and teachers who participated in Cohorts I-III. These rosters would be used as the basis for retrieval of archival records or collection of new data. However, because of the long periods of time that have elapsed since the last data were collected, such studies should be considered only as supplements to studies for which LE data already exist. In 1979-80, the first year that new data collection could be considered, children will be in the grades shown in Table 1. There is no way to collect data for grades 4-7 for any cohort except retrospectively through interviews or archival searches. It is our judgment that studies that rely entirely on such methods would not be sufficiently credible to warrant the inherently high costs of collecting interview data.* Furthermore, there will undoubtedly be severe attrition problems due to the mobility of children in the intervening periods of four to seven years.

Table 1

Grade in 1979-80 by National Evaluation Cohort

Cohort	Grade
I_{EK}	10
I_{EF}	11
II_{EK}	9
II_{EF}	10
III_{EK}	8
III_{EF}	9

4.1.2 Use of the National Evaluation Data Base to Supplement Sponsor-Supported Data Sets

Existing data bases that include LE data might be strengthened by merging them with the national evaluation file to correct specific deficiencies, such as lack of pretests, background, or affective data. In order to assess the usefulness of such data merges, we investigated the details and deficiencies of sponsor and site data sets. A case-by-case analysis of these data sets is provided in Goodrich and St.Pierre (June 1979). In most cases, the overlap of national evaluation and sponsor/site samples was too small to yield a useful merged study sample. In only one case (Philadelphia) did we find both that the national evaluation measures were useful additions and that the sample overlap was adequate.

---

*This judgment is not shared by all of our consultants on this project. Dr. Walter Haney, in particular, feels that small-scale research efforts employing archival and/or retrospective interview data can succeed in producing credible findings, provided that the studies are sufficiently focused on particular questions. Attitudinal or life chances research questions might be candidates for this approach. We would recommend such studies only after the emergence of specific policy issues results in formulation of such specific research questions.

In Philadelphia, the national evaluation sample overlap is nearly complete, and the national evaluation measures would be very useful in analyses. The national evaluation data base would fill in some achievement measures in grades K and 1 now unavailable in the Philadelphia files, and would provide needed socioeconomic and affective data. Furthermore, the national evaluation data could also be used to help restore portions of the Philadelphia files that were accidentally scratched; the alternative would be to re-create the machine-readable files from available hard copy. For further details on this proposed merge, refer to Goodrich and St.Pierre (March 1979) and to Section 5.7 of this report.

4.2 Census of Sponsor-Maintained Data Sets

Our first step in identifying sponsor/site data bases was to conduct a census of sponsors to find all data sets with promising LE data. A list of desired general information about data base structure and content was constructed and used as the basis for informal interviews at the annual FT sponsor conference held at Ypsilanti, Michigan in October 1978. Informal interviews were conducted with all sponsors attending the conference except for Northeastern Illinois University and EDC. These two sponsors, along with two who were not present at the conference (Prentice-Hall and University of California at Santa Cruz), were interviewed later by telephone.

All potential data sets are classified as follows: Table 2 (sponsors with insufficient data), Table 3 (sponsors with site-level data), Table 4 (sponsors with multi-site data), and Table 5 (sites with multi-sponsor data). Those sponsors shown on Table 2 had insufficient data to support LE studies; hence they were excluded from further work. The specific reasons for exclusion are given in Table 2, but no other characteristics of the data are supplied. The remaining ten sponsors had study units with at least 100 subjects and data extending to grade 5. In placing data sets in Tables 3 and 4, we judged that, with the former, only separate site-level studies could be conducted, while with the latter, sponsor-level analyses appear possible. Finally, Table 5 identifies two sites (Philadelphia and New York City) where multiple sponsors were used.

After selection of the sponsors and sites that possessed the most promising LE data sets (Tables 3, 4, and 5), interviews were conducted on site or at sponsor headquarters to obtain more complete information about

Table 2

Sponsors with Insufficient Data

Sponsor	Model	Comment
Clark College	Home-School Partnership Model	No data beyond grade 3.
Education Development Center	EDC Open Education Follow Through Model	No data beyond grade 3.
Hampton Institute	Hampton Institute Nongraded FT Model	Achievement test data for grades 4-5-6 exist (NYC, NJ) for 1978 but sample size is only about 30.
High/Scope Foundation	Cognitively Oriented Curriculum Model	No data beyond grade 3.
University of North Carolina	Parent Education Model	Achievement test data exist only for grade 4 (one site) with sample size of 128. Otherwise, no data beyond grade 3.
University of North Dakota	The New School Approach	An informal reading survey was given to 10-20 children in grades 4-5-6. Otherwise, no data beyond grade 3.
Northeastern Illinois University	Cultural Linguistic Approach	Achievement and Coopersmith data are available for 25-40 students in grades 4-5. Otherwise, no data beyond grade 3.
Prentice-Hall	Prentice-Hall Model	No data beyond grade 3.
University of California at Santa Cruz	Culturally Democratic Learning Environments Model	No data beyond grade 3.

Table 3

Sponsors with Site-Level Data

Sponsor	Model	Sites	Grades	Measures	Machine Readable?
University of Arizona	Tucson Early Education Model	Wichita, KS	1-6	Achievement Tests, Productive Language	No
Bank Street College*	Bank Street College of Education	Boulder, CO Brattleboro, VT Cambridge, MA Elmira, NY Huntsville, AL New Haven, CT New York City, NY Plattsburgh, NY Rochester, NY Tuskegee, AL Wilmington, DL	K-7	Achievement Tests	No
Far West Labs*	Responsive Education Program	Lebanon, NH	1-6	Achievement Tests	No
Georgia State University	Parent Supported Application of the Behavioral Oriented Prescriptive Teaching Approach	Natchitoches Parish, LA	1-6	Achievement Tests	Yes
University of Pittsburgh*	Individualized Early Learning Program	Akron, OH Belcourt, ND Keystone, PA Montevideo, MN Randolph Co., WV Texarkana, AK Waterloo, Iowa	K-6	Achievement Tests	No
Southwest Educational Development Laboratory*	Bilingual/Bicultural Model	Cutler-Orosi, CA Los Angeles, CA San Diego, TX St.Martin Parish, LA Tulare, CA	K-6	Achievement Tests, Spanish Reading	No

---

*Sites are so diverse in terms of level of implementation that separate studies are clearly required.

Table 4

Sponsors with Multi-Site Data

Sponsor	Model	Sites	Grades	Measures	Uniform Across Sites?	Machine Readable?
Fordham University	Interdependent Learning Model	Atlanta, GA New York City, NY	1-6	Achievement Tests, Problem Solving, Phonic Skills, Many Noncognitive Measures	Mostly	Yes
University of Georgia	Mathemagenic Activities Program	Pickens Co., GA McCormick Co., SC Gulfport, MS	1-6	Achievement Tests, Piagetian, Ravens, Cognitive Style, Locus of Control, Language	No	Yes
University of Kansas	Behavior Analysis Approach	New York City, NY Philadelphia, PA Portageville, MO Trenton, NJ Kansas City, MO Louisville, KY Waukegan, IL Meridian, IL N. Cheyenne, Mont. Hopi, AZ	K-5	Achievement Tests	Yes	Yes
University of Oregon	Direct Instruction Model	Flippin, ARK E. St.Louis, IL Smithville, TN Uvalde, TX Dayton, OH Tupelo, MS Flint, MI	K-6	Achievement Tests	Yes	Yes

Table 5

Sites with Multi-Sponsor Data

Site	Sponsors	Grades	Measures	Machine Readable?
Philadelphia Public Schools	Bank Street EDC Kansas North Carolina Philadelphia Process Parent Implementation SEDL	1-8	Achievement Tests, Attendance, Information on competing treatments	Yes*
New York Public Schools**	Kansas Self-Sponsored Bank Street Oregon High/Scope	?	Achievement Tests	?

---

*From 1974-75 on, data are available in computer files, but (because of accidental tape scratching) data from 1971-72, 1972-73, 1973-74 are available only in hard copy. In order to fully utilize the Philadelphia data, it will be necessary to regenerate these files either by punching the records from hard copy or by merging relevant data from the national evaluation data base.

**Information from New York City was received too late for serious consideration in this study.

the data bases. Typically, these interviews involved two staff members from AAI, the FT sponsor, and the person in charge of the data. A loosely structured form was used to organize information and to serve as an informal interview guide. Reports, publications, and any other materials pertaining to the data base were collected. The resulting file of information was used as a source for the data base summaries presented later in this report and those provided in Goodrich and St.Pierre (June 1979).

4.3 Criteria for Selection of Sponsor/Site Data Sets

Our next step was to select data sets suitable for future study by specifying standards to be met. The following standards are used:

(1) The level of analysis can be either the sponsor or the site, provided that the site data can be viewed as arising from the same basic early childhood education model and that sample sizes are sufficiently large.

(2) The base of comparison must be local, in the form of a local comparison group or local norms. Demographic matches must be at least fair to good so that ANCOVA adjustment techniques are credible. Comparison groups can be either *a posteriori* or *a priori*.

(3) Attrition must not exceed 75% from entry to post-FT grades.

(4) There is no requirement that the data set include model-relevant dependent and intervening measures, but special consideration will be given to data sets where model-relevant measures are present. Data must extend at least to grade 6. Well-matched control groups will not be required in cases where sponsor hypotheses about the relation of intervening variables and achievement measures can be tested.

(5) Sample size must be at least 100 (and preferably 250) in FT and NFT groups separately.

(6) Test administration must be up to the standards of professionally supervised, teacher-administered tests.

(7) Measures of post-FT curriculum and process are not required.

In the remainder of this section, we discuss these criteria and their methodological justification.

4.3.1 The Level of Analysis

Possible levels of analysis for LE studies include the school, site, sponsor, or complete aggregation of FT data. Within any of these levels, one can aggregate the entire population of subjects or well-defined sub-populations. There are two major kinds of questions to be addressed in choosing a level of analysis, substantive and instrumental. The substantive questions to be addressed by a particular LE study in large part define the appropriate level of analysis. Instrumentally, we must find the largest regions of the data where a common analytic model can be used and where the population to which results apply is well-defined and relevant to the needs of USOE and other FT stakeholders.* Instrumental and substantive considerations must be taken together in selecting a reasonable level of analysis.

The possibility of conducting LE analyses cutting across all sponsors was ruled out by the diversity of data structures and contextual conditions. The diversity, documented by Goodrich and St.Pierre (June 1979), includes the following:

- The variables selected for measurement differ from site to site and from sponsor to sponsor, although most lie in the domain of academic achievement. Only a few scattered non-achievement measures exist.
- Achievement measures involve different tests and test forms within sponsors and within sites, including the California Achievement Tests, the Comprehensive Tests of Basic Skills, the Iowa Tests of Basic Skills, the Metropolitan Achievement Tests, the Wide Range Achievement Test, and others. Higher levels of of analysis require reduction of test scores to common metrics.
- Test schedules differ by year, grade, and time of year (fall, winter, spring) within and across sites.
- Sites differ as to population characteristics, level of cooperation, degree of model implementation, and many other contextual characteristics.
- The quality and availability of comparison groups or other bases of comparison vary.

Therefore, we identified the largest within-sponsor units (schools, sites, subsets of sites) that qualify for study in terms of sample size, basis of comparison, and other criteria specified in this section, and that are sufficiently uniform in terms of context and data structure. Where there are no qualifying subsets for a given data base, the data set is dropped from further consideration.

---

*One of our consultants on this project, Dr. Leslie D. McLean, has called such a data subset a "reasonable region of inference."

4.3.2 Bases of Comparison

Many credible studies have been based on the use of non-randomized, more or less well-matched local comparison groups, despite the technical superiority of randomized designs. But well-matched local comparison groups are not always available in sponsor and site data bases. Some of the bases of comparison that have been suggested as substitutes for local comparison groups are the following:

- Norming scores relative to some national reference population. These can be expressed in any of several metrics, e.g., standard scores, normal curve equivalent scores, percentiles.
- A national sample of Title I children from the Sustaining Effects Study. (Hoephner et al., 1977)
- Norming scores relative to a local reference population; e.g., city-wide or sub-district averages.
- Unconventional comparison groups. Examples include older and younger siblings; "advance control groups" (i.e., earlier untreated cohorts); and distal control groups (e.g., those from neighboring school districts).

Sponsor analyses have used most of these alternative bases of comparison; but, as Hartman and Acland (1973) note, sponsor and site analyses have been diminished in credibility and usefulness because of uneven standards of methodological rigor and thoroughness. We see the need to set clear minimum standards for selecting bases of comparison for LE studies and have set the standards as high as seems feasible by demanding that local comparison groups be used.

National norming scores are problematic except for the evaluation of models where the clear-cut goal was to achieve growth at "normal" rates. National norms are not generally acceptable as FT comparison scores for several reasons. First, the reference populations used by test publishers can and do differ markedly among publishers and for local populations of interest. Witness the difference between Wide Range Achievement Test and Metropolitan Achievement Test percentiles for the same children as reported, e.g., by Hartman and Acland (1973). Second, the educational effectiveness of any standard curriculum varies widely, not only geographically but across urban/suburban/rural distinctions. It is desirable to determine gains of FT relative to local conditions and local populations, because national norms may mask possible delayed effects (e.g., delayed effects detected with a local comparison group may not be apparent when national norms are used). For instance, consider a pattern in which neutral or negative effects at grade 3 are followed by positive ones in grades 4 through 6. This result is difficult or impossible to interpret without a comparison group. It would be as plausible to say that the post-FT environment is superior to the FT environment as to say that the FT program has a delayed effect. To assure ourselves that we are not looking at a negative FT effect, we need a local comparison group that was NFT but shared the post-FT environment with the FT group.

Other norms groups may, however, prove useful in assessing the delayed effects of FT. An especially interesting norming population is that from the USOE-sponsored Title I Sustaining Effects Study (Hoephner et al., 1977), where approximately 12,000 Title I and 8,000 other compensatory education children were tested in about 125 Title I schools nationwide. Children in grades 1-6 were given the California Tests of Basic Skills in both fall and spring for three years running (1976-1979).

While a group of Title I children is likely to provide a better comparison for FT than a national norms group, we are reluctant to let national or Title I norms be the sole basis of comparison for a sponsor or site. Therefore, we recommend that the use of norms be limited to those sponsors/sites that already have acceptable local comparison groups. In this way, the use of norms will enhance analyses based on reasonably well-matched local comparison groups.

In some instances, sponsor analyses use city-wide or sub-district averages as a base for comparison. Here again, we must worry about the variety of population characteristics and school conditions. Use of these comparison populations does not differ markedly from the use of comparison schools in nearby neighborhoods, a technique that traditionally has been acceptable in educational evaluation. The problem is the extent to which populations and school conditions are well-matched. Local comparisons of this type can be acceptable if the matching is demonstrated to be adequate.

A related base of comparison is the distal control group (e.g., one from a neighboring county). The question that must be addressed here is "How distal can the control group be?" Next county or neighboring school system controls are often subjected to an NFT treatment that is different from the one they would encounter in the FT system. For this reason, we would not accept such comparison groups.

There are also problems with unconventional control groups of various sorts. Older siblings or advance (pretreatment) comparison groups may be initially well-matched in terms of population chracteristics and may be true "NFT"; but to use them as controls is to assume that NFT conditions in, say, 1972 are equivalent to NFT in, say, 1975. One can easily compare successive FT cohorts in existing sponsor analyses or in the national evaluation to determine that cohort effects, i.e., temporal changes, can equal or exceed FT/NFT effects. Changes in site, population, administration of tests, etc., can all contribute to cohort effects. For sibling studies, we must also be concerned about birth-order effects, which are currently the subject of some controversy (Zajonc, 1975; 1976). In short, we believe that advance control groups and sibling control groups cannot support sufficiently definitive evaluation studies.

To recapitulate, we select as suitable for LE studies only those data sets where reasonably well-matched local comparison groups or local norm scores at the neighborhood (or subdistrict) level exist. To assess matching, we rely mainly on sponsor judgments, which have appeared to us to be quite candid and useful. Analyses to determine empirically the degree of FT/NFT match on important pretreatment variables (e.g., pretest score, socioeconomic status) should, of course, be carried out in any LE study.

Although we would prefer comparison groups selected at entry level instead of a posteriori comparison groups, we accept either. A posteriori groups must be well-matched on demographic variables and should have shared recent educational (post-FT) experiences of the FT group. This form of matching can be very important, since differences in post-FT experiences are among the more difficult to cope with.

The use of other bases of comparison (e.g., national norms or Title I norms) is recommended only when local comparison group or local norms are also available. Multiple comparison groups are valuable since they can provide several estimates of treatment effect, which, although each may be biased, can assist in bracketing the true treatment effect.

4.3.3 Attrition

Most of the analyses performed so far by FT sponsors and sites are "quasi-longitudinal"; i.e., they compare cross-sectional averages grade by

37

grade for each cohort. These analyses maximize sample size but are subject to attrition effects. Attrited students often differ from non-attrited students in important ways. For instance, less thriving children from unstable households are likely to be overrepresented in the attrited group. Alternatively, successful FT children may be systematically diverted to regular classrooms. A variety of similar biasing mechanisms exists. These are strong arguments for selecting the longitudinal portion of the sample so that each intertemporal comparison involves the same children. However, a longitudinal sample may represent only a small portion of the original sample, and the results of a longitudinal analysis may not be generalizable to the original population of interest. The adequacy of a longitudinal sample for supporting statistical analyses is easy to check using tables provided in Cohen (1977). But even when statistical power remains adequate, the generalizability issue is a problem. We may, for example, be examining a population composed entirely of FT "thrivers," differing in substantial ways from the attrited group.

It is customary to check the equivalence of attrited and non-attrited groups via t-tests (or similar tests) involving school readiness or any of a variety of demographic measures. These tests are usually not particularly useful except when equivalence is rejected. Failure to reject the null hypothesis can as often be blamed on lack of statistical power as on true absence of a biasing effect. Therefore, we accept such tests as evidence of no difference only when statistical power is demonstrably adequate; i.e., at least .80 to .90 for group differences of practically significant magnitude. This is not often the case in studying the effects of educational programs.

Thus, when attrition is severe, the extracted longitudinal sample is usually inadequate. The only solution is to restrict analysis to data sets where the longitudinal sample size assures adequate statistical power and attrition is not so severe that generalizability seems threatened. We therefore require that at least 25 percent of the sample for LE studies be longitudinal.

4.3.4 Dependent Measures

Some FT sponsors attempt to influence achievement test scores directly; others attempt instead to secure achievement score gains indirectly by influencing domains that are presumed to intervene with or mediate achievement score effects. Examples of these domains are Piagetian cognitive variables (e.g., Georgia), the home environment (e.g., Georgia State), various affective variables (e.g., Bank Street, Far West Labs), or the ability to speak and understand the English language (e.g., SEDL). Comprehensive quantitative studies of post-FT effects for most models will therefore require a wide range of dependent measures spanning affective, cognitive, and process domains, as well as achievement.

USOE's national evaluation was criticized frequently on the narrowness of its measures and the irrelevance of achievement tests as outcomes for some of the models. The few non-achievement measures used (Intellectual Achievement Responsibility Scale, Ravens, Coopersmith) were criticized on grounds of both validity and reliability. At the start of the present project, it was hoped that the sponsors had avoided these faults of the national evaluation by obtaining non-achievement measures through available tests or through tests of their own construction. For the most part, sponsor and site measures proved to be even narrower than those of the national evaluation, although several new and useful instruments were developed. The High/Scope Productive Language Assessment Test, the Bank Street Observation System, the Atlanta Public Schools Basecheck, and several measures of model implementation are among these.

It became clear quite early in our project that, if we were to require that a sponsor's data set include measures of the most important outcome variables according to their model, then we would have to eliminate most data sets. In fact, only two (Georgia and Fordham) possess extensive model-relevant data of this type. However, where these measures are present, we would relax other standards discussed in this section, provided that analyses could hope to validate or invalidate basic sponsor hypotheses concerning later effects.

LE studies based on existing data will concentrate on achievement test scores, simply because this type of data is most readily available. Opinions of the sponsors on this projected narrowness were solicited at a conference held at AAI in March 1979. There was near unanimity among the ten invited sponsors* on the appropriateness of achievement test scores for LE evaluations. The sponsor nearest to dissent (Bank Street) felt that, while achievement test scores were irrelevant in themselves, they were a part of post-FT school reality, and doing well on tests was a valuable coping skill relevant to life chances. Regardless of whether their immediate goals lie in the achievement domain, all attending sponsors felt that their ultimate post-FT goals did include achievement test performance. They recognized that, rightly or wrongly, most public school systems currently see the teaching of basic skills as their most important mission, that the tests reflect basic skills, and that, if FT is to serve students in the public schools, its sponsors must acknowledge this.

---

*Only those sponsors having some useful LE data were invited.

Although there is acceptance of achievement test scores by the ten sponsors who attended the meeting, these sponsors included only those with LE data. Those who did not collect such data on their own or attempt to assemble a data base from site-level data are more likely to dissent. Thus, we do not expect unanimity across all sponsors.

Two other domains of dependent measures are relevant to LE studies. The first are the "life chances" variables of attendance, grade retention, special education placement, drop-out rates, etc. Where these data can be obtained--and they are often available in site files if not in sponsor files--it is recommended that they be added to the LE data file and used as dependent measures.

The second domain to be considered consists of interview data that might be collected from parents, teachers, and children. Such data could concentrate on retrospective assessments of the FT to post-FT transition process. Also of interest would be post-FT opinions and recollections about the program. Because of the expense involved, we do not recommend large-scale data collection of this type; rather we judge that one or two such efforts should be concentrated in sites where existing achievement data are already quite strong, so that descriptive (soft) studies can be used to support quantitative (harder) studies.

4.3.5 Statistical Power

The purpose of this feasibility study is to identify data bases that allow potentially useful statistical tests of post-FT effects. Our expectations about the size and direction of effects are, to a large extent, a function of the effects found in previous examinations of FT data. First, the national evaluation did not find many large positive effects at the end of the relatively long three-to-four-year treatment. Based on this, the expectation of positive effects that increase in size during the post-treatment years may not be reasonable. Second, of the many existing sponsor and site studies, very few show strong indications of sleeper effects or sustained normal gain rates (where earlier normal gains were found).

Statistically, then, FT's history is one of a great many "null" effects rather than striking positive effects. In light of these findings, failure to reject the null hypothesis often has been wrongly interpreted as statistical proof that the null hypothesis is true: practical significance and statistical significance are not distinguished in the minds of many readers. Consequently, we recommend that no study unit (sponsor or site, as the case may be) be considered unless the total sample size is at least 100 in FT and NFT each. This requirement was arrived at as follows. A common definition of the minimum educationally significant effect size is .25 of the child-to-child standard deviation of the outcome measure. The most common level of significance used in hypothesis testing is .05; i.e., we wish to reject the null hypothesis by chance alone no more than .05 of the time. The _a priori_ probability that an effect of .25 SD is incorrectly declared null is .41 at a sample size of 100 (t-test at alpha = .05; see Cohen [1977]). Thus, sample sizes of less than 100 yield hypothesis tests of abysmally low statistical power and should not be considered. Only for samples as large as 250 does power approach a "comfortable" level, i.e., a type II error rate of .20 or less. We reluctantly accept samples of about 100 but would prefer much larger samples. To the extent that subsamples of children within a site are of practical interest, the data will have to be broken down, and the sample size required would be 100 per subpopulation of children, rather than a total of 100 per site.

We also recommend that, when studies are conducted and presented, each null finding be accompanied by a statement about the power of the test in much the same way that significance levels are conventionally presented.

4.3.6 Data Quality

The data now assembled in sponsor and site data bases have been collected by many different methods. Test administration and scoring standards vary from extremely rigorous to sloppy. For the most part, one cannot now make highly objective assessments of the quality of the data and must instead rely on the subjective opinions of sponsor and site personnel. We have solicited these opinions from all sponsors. Replies have been frank and useful, and in several cases data subsets have been determined to be of low quality. However, in no case has an entire data set been judged unusable for this reason alone. Except in isolated cases, data are of the routine quality to be expected from teacher-administered tests under professional supervision. Data sets must reach this subjective standard in order to be considered in our study.

There are also problems with the validity of some measures. Most FT analysts know that the affective measures in the national evaluation were severely criticized and that this hurt the credibility of the study somewhat. Several measures considered for LE studies are also controversial or of unproven validity. The test most often criticized is the Wide Range Achievement Test, on the grounds that, in the later grades (4-6), it is non-conceptual and even trivial. We comment on the validity of specific measures in our discussion of specific LE studies under consideration.

4.3.7 Measures of Post-FT Curriculum and Process

An extremely serious problem in the analysis of post-FT data is the dependence of achievement scores on the most recent educational events in the children's history. It is reasonable to expect that post-FT achievement is a stronger function of recent treatment and process in standard classrooms than of the more removed FT treatment. Yet children are dispersed through so many different classrooms, schools, and neighborhoods that it is virtually impossible to control for these post-FT conditions. If ignored, these variables may alter post-FT outcomes in a seemingly random way unless coherent groups of children are exposed to the same post-FT conditions. Thus, we expect the result to be decreased visibility of the FT treatment because of new, uncontrolled variance. In effect, FT outcomes may be attenuated as fractions of total variance, and statistical power is decreased.

We see no satisfactory post hoc method of dealing with this problem. Perhaps, given closer examination of some of the data, it will be possible to measure some post-FT inputs in some sites and, if possible, these should be introduced into analyses. In selected cases, an interview-based attitudinal study as discussed above can be used to begin study of the variety of post-FT environments and transitional effects. If future new experimentation is performed, post-FT treatment should be considered as an important design variable.

For now, the availability of information on post-FT process variables is not an important criterion in selecting sponsor and site data bases for analyses.

4.4 Qualifying Data Sets

Goodrich and St.Pierre (June 1979) provide a detailed description of each of the data bases identified in Tables 3, 4, and 5. These descriptions are summarized in Table 6. The selection criteria presented in Section 4.3 were applied and six sponsor and/or site data bases were identified as being suitable for study. Here we summarize our recommendations.

Two of the sponsor/site data bases (Oregon and Fordham) afford especially good opportunities, since they surpass considerably the minimum methodological standards (ample sample size, availability of a wide range of measures, good local control groups), and because the results might be generalizable. Along with the Philadelphia data base recommended for study by Goodrich and St.Pierre (March 1979), we have a total of three top-ranked data bases. These are as follows:

1. Philadelphia, merged with the national evaluation data base, concentrating on Cohorts I-III;
2. Oregon, concentrating on five sites with well-matched local control groups and adequate sample sizes; and
3. Fordham, including both Atlanta, GA and New York City, NY.

The remaining three data bases that meet our criteria are weaker methodologically and/or in generalizability. They focus on single sites with marginal sample sizes. These are:

4. Arizona (Wichita, KS);
5. Georgia State (Natchitoches Parish, LA); and
6. SEDL (San Diego, TX).

Any one of the latter three data bases would support smaller studies that, since they are restricted to single sites, would not allow general conclusions to be drawn about the educational models involved. However, if combined with the other studies and with existing work on post-FT effects, they may help

Table 6

Summary Assessment of Data Bases

Data Base	Measures	Comparison Groups	Estimated Longitudinal Sample Size	Recommendation
Philadelphia Public Schools	Uniform achievement data for several sponsors.	Cross-sponsor comparisons. Well-matched local comparison groups to 1975.	Approximately 100 per sponsor per cohort.	Multi-sponsor study.
Fordham/ Atlanta, New York City	Achievement, cognitive, affective. Achievement data are longitudinally uniform.	Well-matched local comparison groups in Atlanta, fair to poor in New York City.	1,000s (Atlanta). 100 (New York City).	Study two sites jointly.
Oregon	Uniform achievement data across several sites.	Local comparison groups in some sites.	100s.	Study five sites with local comparison groups (sponsor level) and adequate sample size.
Arizona/ Wichita	Uniform longitudinal achievement data.	Randomized across treatment/control groups. There are also non-random FT classmate, low income groups.	40 (of initial 100) longitudinal children per cohort.	Site-level study
Georgia State/ Natchitoches	Longitudinally uniform achievement data.	Well-matched local comparison groups for 3 of 6 schools.	Low 100s across all cohorts.	Site-level study.
SEDL/ San Diego	Non-uniform achievement data. Some longidudinal uniformity.	Adequately matched local comparison groups.	75-150 per site for each cohort.	Site-level study.

Table 6 (cont.)

Summary Assessment of Data Bases

Data Base	Measures	Comparison Groups	Estimated Longitudinal Sample Size	Recommendation
Bank Street	Non-uniform achievement data across several sites. Some sites have longitudinal uniform achievement tests.	Scattered and poorly matched local comparison groups.	100s in larger sites.	No study.
Far West	Achievement.	Good NHSFT/NFT match, poor HSFT/NFT match in Lebanon. No comparison group in Goldsboro.	Low 100s.	No study.
Georgia	Mostly uniform achievement, Piagetian, affective data.	Poorly matched local comparison groups.	Low 100s with full data.	No study.
Kansas	Uniform achievement data across several sites. K-3 consumer satisfaction.	Well-matched local comparison groups in urban sites only.	100s for each of 3 cohorts	No study.
Pittsburgh	Achievement data, some of it uniform.	"Advance control" group only.	Very small longitudinal samples. Crossectional in 100s.	No study.

substantially in fitting together a more complete picture of FT. They would be useful mainly as important quantitative case studies. Note that, if the Los Angeles data from the SEDL model become available (see Section 5.5), our recommendation of a SEDL study would be much stronger.

There are two sponsors where one or more data sets might support studies but where more information is needed. These are as follows:

1. Pittsburgh. The existence of useful comparison groups and useful within-site sample sizes would have to be demonstrated in their ongoing data inventory.
2. Far West Labs, St. Louis only, if the data prove to be available in pupil-by-pupil records.

A data base recommended earlier by Goodrich and St.Pierre (June 1979) is supported by the University of Georgia. Although sample sizes are marginal, matched control groups are unavailable, and measures are non-uniform across cohorts and sites, the data include an interesting and broad panel of affective and Piagetian measures. The basis of our earlier recommendation was that the data might support interesting panel analyses and allow the testing of certain causal hypotheses. The results might be of substantial interest to educational psychologists. In investigating these data further, we sought to construct analyses that would help to illuminate specific questions about later effects of FT. It was determined that the sample size available to test any specific hypothesis was too small unless data were pooled across cohorts. This is not feasible because the panels of Piagetian measures were different each year. We have therefore decided not to recommend a study of these data. The data are a gold mine for scientifically-oriented theses and dissertations (witness the excellence of Georgia's report series), but not for policy research.

The Bank Street data do not meet the study criteria, especially in terms of appropriate bases of comparison and adequate sample sizes. Findings from site-level analyses conducted in Hawaii, Fall River, and New Haven could be integrated with those from the separately recommended Philadelphia study in an attempt to form general conclusions about post-FT effects of the Bank Street model.

We do not recommend study of the Kansas data, largely because the data extend only through grade 5--too short a horizon--and because of the weakness of the Wide Range Achievement Test in the upper elementary grades. The existing analyses conducted by Kansas are of good quality and directly address the problem of fade-out. These analyses should be used in any general attempt to integrate and interpret post-FT effects across sponsors.

The collection of limited new data in the form of interviews with teachers, children, and parents would be useful in certain cases. This effort might be concentrated in a few sites, e.g., in Atlanta and New York City (the two Fordham sites), where a variety of affective and other non-achievement measures already exist. The largely descriptive analysis of these interview data would be intepreted in light of the Fordham measures. Interest would focus on FT fourth grade transition problems and on changed attitudes.

5.0 DISCUSSION OF DATA BASES SUITABLE FOR LATER EFFECTS STUDIES

Six data bases have been identified as meeting our criteria. This section explores them in some detail. For each data set, we first give a brief description of the sponsor's educational model and hypotheses for post-FT effects. Then we provide a general description of the data base. Finally, we describe the studies that the data can support. This includes discussions of research questions, analytic strategies, and cost estimates. The general research questions identified in the first section of this report are further specified in the discussion of each data base.

5.1 General Analytic Approach

In this section, we present a general discussion of issues common to all or several of the six studies, so that individual study plans can be made more succinct. Each plan is divided into three analytic phases, to begin when data have been reduced to clean, longitudinal child-level records and are available for analysis. These stages are preliminary exploratory analyses to establish hypotheses and detailed analytic strategies, the conduct of the summative analysis, and the validation of conclusions.

5.1.1 Preliminary Exploratory Analyses

Preliminary analyses involve seven areas of inquiry:

(1) determination of the standard of comparison and required intergroup adjustments;

(2) determination of the level of analysis and unit of analysis;

(3) selection of longitudinal or quasi-longitudinal samples;

(4) assessment of psychometric properties of measures;

(5) conduct of exploratory data analyses;

(6) construction of analytic data base; and

(7) assessment of generalizability.

Determination of the Standard of Comparison and Required Intergroup Adjustments

Each data base involves one or more local comparison groups that serve to establish NFT levels of outcome in the quasi-experimental paradigm. Except in the Wichita study, local comparison groups are not randomly selected and are imperfectly matched with the FT treatment group. Differences between

FT and NFT groups must be studied (descriptively) in the detail allowed by available data, and an appropriate adjustment strategy must be selected.

Demographic variables, preschool experience, and other initial status variables should be studied descriptively and the results interpreted in light of the selection criteria. All data sets that meet our criteria include adequate local comparison groups, but, in some cases, the comparability of certain cohorts is defective because of disruptive uncontrolled events. The product of the examination should include selection of cohorts and sites where credible comparisons are feasible. Defective data subsets should be discarded.

Adjustment strategies that should be considered to account for different FT/NFT initial status include ANCOVA with or without disattenuation, LISREL analysis, and Value-Added analysis.

The ANCOVA approach to intergroup adjustment is the one most frequently used, although the method has several defects that might lead to biased estimation of treatment effects. Most of the problems stem from specification errors of various sorts. The most frequently discussed problem with ANCOVA is that of the fallible covariate (Porter 1967, Lord 1960). When the "true" model involves simple linear dependence on a "true" covariate like socioeconomic status, but the covariate can be measured only imperfectly, the standard ANCOVA approach yields a treatment effect estimate biased against the "less able" group. This is the basis of the argument for disattenuation via the Lord-Porter correction.

Cronbach et al. (1977) have showed that, in practice, disattenuation can actually increase the treatment effect estimate bias. This occurs because the structure of the "true" model is in practice unknowable, and specification error is always present in one form or another. Assumptions of linearity, homogeneity of regression, and the inclusion of all relevant independent variables are rarely valid and, moreover, are very difficult to test statistically. Tests for heterogeneity of regression, for instance, are usually so lacking in statistical power that failure to reject the null hypothesis means little or nothing. The ANCOVA procedure, especially with disattenuation, requires willing suspension of disbelief.

The great advantage of classical randomized controlled experiments is that the estimation of a treatment effect does not depend on assumptions about an underlying true model. In social science research, we can rarely use such an approach, but we can at least try to minimize the role of untestable assumptions. We therefore recommend that traditional ANCOVA be accompanied by a new ANCOVA approach, one based on the theory of linear discriminant functions (LDF), which appears to require a minimum of assumptions about the true model.

Cronbach et al. (1977) showed that ANCOVA estimates are biased unless the assumed covariate X is either "true" or proportional to the "complete" linear discriminant function. Their proof of unbiasedness in the latter case extended Goldberger's (1972) result that, if assignment to treatment and control groups is carried out on the basis of a cutting value on a measured variable (e.g., income), then regression on that variable and the treatment variable results in unbiased treatment effect estimates. The result holds true whether or not the variable is measured with error, and is independent of the relation between the measured variable and the variables that occur in a "true" model. There are, in other words, no strong assumptions about the relation of the covariable to the process under investigation.

Appendix 1 develops Cronbach's and Goldberger's discoveries further in the form of the linear discriminant function covariable method. In its simplest form, the linear discriminant function based on a stock of available covariables is pre-computed and used as a univariate covariable. Under reasonable assumptions, the resulting treatment effect estimates are unbiased, even though the linear discriminant function may bear little relation to the unknowable "true" covariable. Although this approach is new, we recommend using it in conjunction with traditional ANCOVA because so little is assumed in its formulation.

In the LISREL approach (Magidson 1977), it is assumed that equation (1) holds (see Appendix 1) but that the covariate X is not directly observable. Measurements of mother's education, family income, etc. are taken to be error-corrupted measures of X. The program LISREL is used to obtain maximum-likelihood estimates of regression coefficients and the treatment effect. These estimates are automatically "disattenuated" for the presence of measurement error, without requiring a priori knowledge of the reliability of measurements.

We recommend against such an approach for LE studies because it depends strongly on the existence of a true model of the type that must be postulated, and because of the computer expenses involved.

Value-Added techniques were first suggested for studies of cognitive growth by Bryk and Weisberg (1977). One begins by preparing estimates of expected growth through cross-sectional analyses of growth by age before treatment is initiated. These estimates are compared to actual growth in the presence of treatment to calculate treatment effects.

In general, value-added techniques do not apply to the proposed studies because of the unavailability of pre-treatment data over the relevant age range.

Determination of Level of Analysis and Unit of Analysis

Because each child is exposed to a succession of different class groupings over the six- to seven-year period (K through grade 6) of greatest analytic interest, no single class grouping presents itself as a candidate for aggregation. Practically speaking, the unit of analysis must be the child. It is hoped that serial correlations caused by common class effects will be dissipated as the child is exposed to a long sequence of educational environments. If not, the validity of conventional child-level hypothesis testing may be threatened.

Questions that should be asked include the following: What are the significant sources of variation in post-FT outcome measures? What is their relative quantitative importance? What are the implications for study design? The relative contributions of site and school effects to variation in post-FT outcome measures are important because of the necessity of understanding the components in the hierarchical design--child within class within school within site within sponsor. This question of relative importance can be answered directly via variance components analyses (Graybill 1961), which allow one to ascertain the degree to which child-level hypothesis testing is compromised.

Variance components analyses can provide estimates of the actual outcome variance contributed by these several sources, and allow the researcher to calculate the degree to which the summative analyses will be weakened by unwanted variance due to post-FT educational treatments. Treatment variables from preschool to grade 3 that do not contribute substantially to outcome variance can be dropped from the list of sources to be investigated.

Variance components analysis allows estimation of the threat to hypothesis testing posed by intraclass correlations. If FT and post-FT school/classroom variance components turn out to be minimal, then child-level analyses are appropriate. If either FT or post-FT school/classroom effects are strong, one would ordinarily aggregate to the level of the stronger effect. The number of classes or schools, however, is likely to be insufficient for such aggregated analyses. If, as seems likely, there are strong school or class effects, then the effect on hypothesis testing should be investigated. Under these conditions conventional hypothesis tests are biased, resulting in too frequent rejection of the null hypothesis (Singer and Goodrich 1979). A cumbersome but feasible strategy is the use of error components regression models (Nerlove 1971). If necessary, key findings should be examined to determine their sensitivity to intraclass correlation effects.

Caution is required in the interpretation of variance components analyses because estimates are often unstable and may vary widely across groups. This difficulty is inherent in the models and estimation methods used. Conventional expected mean-square methods often result in negative variance estimates, and any important use of variance components analyses should be made only if reasonable confidence can be placed in the stability of estimates.

The determination of a level of analysis must also be made. Several of the data bases include multiple sites, and the question is whether they should be studied in one pooled sponsor-level analysis or whether site-by-site analyses should be conducted. This issue can be settled empirically. Sites (and perhaps cohorts) should be pooled if the same statistical models apply to all sites, and if the substantive research questions make this sensible. This can be determined by statistical tests for homogeneity of regression (Chow 1960) and by the use of descriptive studies of site characteristics (including anecdotal evidence).

Selection of Longitudinal or Quasi-Longitudinal Samples

Longitudinal samples should be selected wherever feasible--simply because results are more persuasive and because individual time series are of interest.

Assessment of Psychometric Properties of Measures

It is important that the reliability and validity of each dependent measure be known. Opinions about the validity of measures and their credibility to the FT research audience must be solicited. We recommend a short but concentrated psychometric analysis of all tests.* The analysis would include factor analyses, intercorrelation analyses, computation of alpha reliabilities, and perhaps some generalizability analyses.

---

*Test reliabilities are not properties solely of the tests, as commonly assumed but, rather, simultaneous characterizations of the test and the population.

Conduct of Exploratory Data Analyses

Exploratory data analysis is always performed, at least informally, as analysts "get their hands dirty" with the data. We recommend a more thorough exploratory phase than is usually carried out, however, using specific new tools such as box plots and stem-and-leaf displays, as well as older techniques like the scattergram. The modern approach to exploratory data analysis is described by Tukey (1977) and by Mosteller and Tukey (1977). The outcome of the exploratory phase of data analysis will be a set of plausible and useful working hypotheses to be confirmed--or disproved--in later, more model-oriented analyses.

Construction of Analytic Data Base

The definition of an analytic data base for summative analyses should be made on the basis of findings in the inquiries noted earlier, particularly conclusions reached in the exploratory data analysis. Data anomalies, specific knowledge that contraindicates use of parts of the data, missing data problems, etc., must be considered. Although many of the decisions must be somewhat subjective, the issues involved are standard within data analysis.

Assessment of Generalizability

In each proposed study, there are limits to the generalizability of the results. The universe of generalizability cannot be determined through statistical methods for most of the studies because of the lack of multiple data sets in different classes. For example, one might ask whether results from Natchitoches Parish generalize to the entire rural South, but since there is no measure of variance within the rural South, there is no statistical method to assess generalizability. On the other hand, the results from the several Oregon sites might make it possible to determine site-level generalizability.

In general, one must base claims of generalizability on reasoning processes and anecdotal data rather than on the statistical treatment of numbers. It is recommended that for each study a statement about the universe of generalizability be made, even though verbal arguments must be cited in lieu of statistical ones.

5.1.2 Summative Analyses

Each of the data sets recommended for study consists of panels of data beginning in grade K (or 1) and continuing longitudinally for several years thereafter. They will support, to greater or lesser extents, a cumulative panel data approach to be explained here. The individual study plans described in Sections 5.2 through 5.7 discuss adaptation of the approach to specific data sets. Table 9 (in Section 5.2.3) exhibits the data structure for a typical cohort (Atlanta Cohort IV) in the proposed Fordham study, where variables are organized into demographic, process, treatment, and test categories. This data structure serves as an example for the following discussion.

The basic theoretical premise is that incremental growth in a measure (e.g., the Iowa Test of Basic Skills for grade 4) depends cumulatively on concurrent learning experiences and all previous experiences. Previous experiences not only supply the child with an inventory of specific skills and items of knowledge, but also with general cognitive, social, and problem-solving skills, and with affective development. Outcomes from the past become inputs to the present process through such changes in the intellectual and affective status of the child.

As the child grows, variance in current outcome measures attributable to specific earlier educational events is expected to decrease, as these earlier events are layered over by more recent educational experience. Events that markedly affect the child's cognitive strategies (as, for example, Bank Street's curriculum is intended to do) might be exceptions to this general expectation; i.e., their influence is hypothesized to persist or even grow. In any case, it is important to understand the process by which effects are propagated through time. Inquiry of this nature has an important history from earlier studies (e.g., the persistence of Head Start effects into the early school years), and the subject is controversial. But, if it is possible to understand and describe this process, the research community will have gained important new knowledge about persistence of educational interventions.

The overall research goal is to construct a cumulative panel model of this growth process, where emphasis is on the causal influence of past and concurrent variables on post-FT performance. Although this ambitious objective may not be fully accomplished, some of the data sets (particularly Fordham) afford a rare and relevant opportunity to apply state-of-the-art techniques in causal modeling (Joreskog 1977, 1978a; Duncan 1975; Boardman and Murnane 1979) and stochastic modeling (Goodrich and Singer 1978). More conventional (and more easily communicated) methods should be used to parallel causal modeling. The following pages present a formulation of this approach.

Our suggested approach to the analysis is similar to that recently discussed by Boardman and Murnane (1979). While they note that their methods are exploratory and tentative, they do offer a promising approach. The goal is to construct a sequence of causal models that relate the data panel for each grade (K-6) to current independent variables and previous data panels. The overall model is identified in steps, beginning at grade K and proceeding grade by grade to grade 6 or beyond. Because of technical problems (multicollinearity, identifiability), parts of the overall model may not be completely determined. However, the effort is made worthwhile by the knowledge gained in the process. We briefly explain the technique.

Let lower-case k represent grade level (k = K, 1, 2, ...), where upper-case K indicates kindergarten as usual. Represent the panel of dependent measures (tests, process) available for grade k by X_k and the independent measures (treatment, demographic variables) by Z_k. Then assume the model:

$$A_k X_k = \sum_{j < k} B_{kj} Z_j + \sum_{j < k} C_{kj} X_j + e_k$$

where A_k, B_{kj}, C_{kj} are matrices of appropriate dimension and e_k is a vector of uncorrelated errors. This model is not identifiable without assumptions about directions of causality. However, when A_k is non-singular, one can left multiply by A_k^{-1} to obtain the so-called (by econometricians) reduced form:

$$X_k = \sum_{j < k} F_{kj} Z_j + \sum_{j < k} G_{kj} X_j + f_k$$

where $F_{kj} = A_k^{-1} B_{kj}$, and $G_{kj} = A_k^{-1} C_{kj}$

and the errors, f_k, are correlated across the data panel; i.e., their covariance matrix Σ_f is not diagonal. The parameters (F, G, Σ_f) of this model can be identified approximately by ordinary regression methods and refined via LISREL estimation. The matrices F and G contain information about persistence of effects directly relevant to our study questions. We expect the size of elements in the matrices $F_{k,k}$, $F_{k,k-1}$, $F_{k,k-2}$ ... to diminish until all elements are negligible; i.e., as a treatment recedes into the past, its observable effect is expected to diminish relative to other effects.

When measures are sufficiently precise and sample size is adequate, it is possible to determine direction of causality among variables where concurrent time series are available (e.g., the Phonic Skill Test, the Iowa Tests of Basic Skills, and process measures given in Table 9). Tests for the existence of feedback and direction of causality were first discussed in the econometrics literature by Granger (1963, 1969) and Sims (1972). Early applications dealt with the relation between money and income (Feige and Pearce 1974, Dy Reyes 1974, Wall 1974) and between advertising expenditures and sales. The formulations of Sims and Granger were unified and extended in a rigorous mathematical formulation by Caines (1976). These ideas were extended to the case of cross-sectional finite-length time series by Caines, Goodrich, and Carlson (unpublished). Carlson (1979) applied these techniques to determine the direction of causality between prison populations and prison capacity. Although the information about causal direction obtained by these techniques is weaker than could be obtained experimentally, it is far stronger than could be obtained without using time series methods.

5.1.3 Validation of Conclusions

If resources permit, formal validation and tests for sensitivity to particular data subsets should be performed. We suggest three possibilities for such examination of the results--examination of residuals, the jackknife, and robust regression. These should be supported by more subjective methods as well; e.g., by the use of scattergrams.

The need to examine residuals for patterns of heteroskedasticity, serial correlations, and systematic behavior is well known (Draper and Smith 1966). To these conventional analyses we would add only the computation of intraclass (or school) correlations of the residuals. Since analyses must (for practical reasons discussed earlier) be conducted at the child level, the intraclass correlation can be used to evaluate validity of child-level hypothesis testing. There may be a need for approximate degrees of freedom corrections using, for example, the method of Greenhouse and Geisser (1959). It may also be feasible to perform error-components regressions (Nerlove, 1971).

Recently, much attention has been paid to so-called robust statistics, especially the jackknife method developed by Tukey (1958) and explained in detail by Mosteller and Tukey (1977). Since the main analytic tool proposed for post-FT studies is regression of various forms, the jackknife should be applied mainly to estimates of regression coefficients and construction of robust hypothesis tests. This will be accomplished by using natural subdivisions of the data--cohorts, schools, classes--as jackknife groups. Estimates of regression coefficients are made for the reduced body of data obtained by excluding each group in turn, as well as for the complete data set. For the estimate excluding group j, a "pseudo-value" for the coefficient b, which we will denote as b^*_j, is obtained via the formula:

$$b^*_j = Kb_0 - (K-1)b_j$$

where b_0 is the ordinary least squares estimate based on the whole data set, b_j is the ordinary least squares estimate based on the subset of data excluding group j, and K is the number of data groups (e.g., classes). The jackknife estimate of b is then given by:

$$b^* = \frac{1}{K} \sum_j b^*_j$$

and its variance by:

$$S^2_* = \frac{1}{K} S^2$$

where S^2 is the sample variance of the pseudo-values. Empirically, these estimates are known to be robust and to afford safe and conservative tests for statistical significance, with greatly reduced sensitivity to quirky data subsets. The technique can be expensive because of the number of estimates required and should therefore be used only to validate the most important conclusions. Computational techniques that greatly reduce the expense are given by Belsley, Kuh, and Welsch (forthcoming), who also present a variety of diagnostic tools that should be used in judging the validity and sensitivity of regression models.

Robust regression techniques are also discussed by Mosteller and Tukey (1977). These techniques reduce sensitivity to particular data points by adaptively downweighting points far from the regression surface. Computer programs exist for calculation of weights and coefficient estimates.

At this point, we proceed to a discussion of the recommended studies.

5.2 Fordham University Data Base

5.2.1 Educational Model and Later Effects Hypotheses

The Interdependent Learning Model is sponsored by Fordham University and uses transactional instruction games, role playing, and programmed instruction as vehicles for teaching in an open, non-authoritarian, non-punitive environment. The direct goals are to foster child cooperation, independence, interdependence, positive self-concept, self-control, self-reliance, verbal fluency, systematic reasoning and problem solving, and self-evaluation, i.e., to foster affective and cognitive growth. These intervening variables are hypothesized to lead to academic competence. Positive effects on achievement are expected in grades K-3, followed by a possible trough (transitional adjustment problems) in grade 4 and enhanced achievement in grades 5-6. Since conventional fourth grade is more competitive than FT, we expect cooperation and independence to wane, while decision-making and problem-solving abilities should be retained.

5.2.2 General Description of the Data Base

The model is implemented on a wide scale in Atlanta and in only one school in New York City. A great variety of affective, cognitive, and achievement data have been collected in both sites. Extensive data from New York are available in a clean, well-organized longitudinal file maintained at Fordham. The Atlanta data reside in two places: a longitudinal file maintained at Fordham, in need of cleaning and restructuring; and in the Atlanta Public Schools files. In order to use them for LE studies, the two files would have to be merged.

Measures Available

A wide array of tests were administered by Fordham, including measures of school readiness (Basecheck, Preschool Inventory, Metropolitan Readiness Test); academic performance (Metropolitan Achievement Test, Iowa Test of Basic Skills); behavioral maturity, phonic skills, test anxiety, general anxiety, and self-esteem (Coopersmith), and intellectual achievement responsibility. These data are available in a complex pattern across grades and cohorts, detailed in Tables 7 (New York) and 8 (Atlanta). The Fordham data can be supplemented by Metropolitan Achievement Test (to 1973-74) and Iowa Test of Basic Skills (from 1973-74) measures administered by the Atlanta Public School system in each April of grades 1-7. The Atlanta Public Schools has also collected state-mandated criterion reference test data, the Boehm Test of Basic Concepts, and the Comprehensive Instructional Program Reading Diagnostic Test. Implementation data are available for every classroom in both sites. Grade retention and special education placement data can also be obtained.

Comparison Groups

In Atlanta, there are five (six before 1976-77) well-matched NFT schools in the same communities as FT schools.

The New York public schools did not permit public identification of comparison schools used in the national evaluation, so Fordham chose its own well-matched comparison school in the block next to the FT school. It is known to not be the same as the national evaluation comparison school.

Sample Sizes

Sample sizes for New York and Atlanta are shown in Tables 7 and 8, respectively. If the Atlanta file is restructured and cleaned, sample sizes will increase by a factor of three to four. In the present Fordham-maintained file of Atlanta data, "dirty" records were discarded rather than cleaned.

Demographic Data

A number of background and experience variables are available, including sex, entering grade, birthdate, ethnicity, pre-kingergarten experience, and the number of months in FT each year. In addition, family, health, and social services information is available for each child.

Test Administration and Validity

Original data were collected from mixed sources. In Atlanta, much of the data was collected by school testing teams supervised by an evaluation specialist. Some other data were obtained from the national evaluation data base at the time of original collection. Data were transcribed at the site

Table 7

Sample Size by Measure, Cohort, Grade, and Test Point for the Fordham (New York) Data Base

		COHORT 0		COHORT 1		COHORT 2		COHORT 3		COHORT 4		COHORT 5		COHORT 6		COHORT 7		COHORT 8		COHORT 9	
MEASURE	GRADE	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Basecheck	K												50	56	53	67	**	**	**	**	**
BMS	K											93		155							
	1									112		146									
	2							89		141											
	3					132		153													
PST	K											90	85	79	67	63	63	**	**	**	**
	1									110	97	110	81	87	80	**	**	**	**		
	2							104	98	95	89	87	78	67	64	**	**				
	3					112	108	88	78	87	93	66	62	**	**						
	4			96	92	93	90	74	86	15	72	**	**								
	5	82	74	74	53	94	100	--	70	**	**										
TASC	1										90		98		**		**				
	2								91		85		71		66						
	3						90		80		82		60								
	4				84		82		78		74										
	5		68		70		92		80												

(continues)

**Data are available and can be added to file.

Table 7 (cont.)

Sample Size by Measure, Cohort, Grade and Test Point for the Fordham (New York) Data Base

		COHORT 0		COHORT 1		COHORT 2		COHORT 3		COHORT 4		COHORT 5		COHORT 6		COHORT 7		COHORT 8		COHORT 9	
MEASURE	GRADE	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
GASC	1										89		98		**		**				
	2								90		85		--		65						
	3						90		79		--		60								
	4				84		82		--		74										
	5		69		73		--		83												
Gilmore	4											56									
	5									68											
	6							64													
Coopersmith								82		89		57									
IARS	3							90		88		57									
ITBS	1															84		52			
	2													72		61					
	3											70		57							
	4											60									
COI								90													

(continues)

Table 7 (cont.)

Sample Size by Measure, Cohort, Grade and Test Point for the Fordham (New York) Data Base

		COHORT 0		COHORT 1		COHORT 2		COHORT 3		COHORT 4		COHORT 5		COHORT 6		COHORT 7		COHORT 8		COHORT 9	
MEASURE	GRADE	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT
MAT-SRI**	3							72													
MAT-NYC*	2			140		125		103	113												
	3	131		107		124	127														
	4	113		136	118																
	5	117	114																		
NYC* Reading	2									93	81	86	75	67	57	56	71				
	3							106	90	92	75	69	76	59	60						
	4					113	95	92	90	72	82	57	77								
	5					109	87	83	77	64	78										
	6							61	--												
SSRA	2															64	31				
	3													59	30						
	4											59	31								
	5									68	32										
PEPSI	2															64	31				
	3													58	30						
	4											59	31								
	5									69	32										

**Indicates data that currently exist in raw form only

*Other MAT data from SRI are available in computer form for years 1971-72, 1972-73, 1973-74, 1974-75

(continues)

Table 7 (cont.)

Sample Size by Measure, Cohort, Grade and Test Point for the Fordham (New York) Data Base

		COHORT 0		COHORT 1		COHORT 2		COHORT 3		COHORT 4		COHORT 5		COHORT 6		COHORT 7		COHORT 8		COHORT 9	
MEASURE	GRADE	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT
PEP*	3									95	76										
	3											73	72								
	3													60	60						
	6							62	77												
OPT*	2															30	30				

*Indicates data that currently exist in raw form only.

(concluded)

Table 8

Sample Size by Measure, Cohort, Grade, and Treatment for the Fordham (Atlanta) Data Base

		COHORT 0		COHORT 1		COHORT 2		COHORT 3		COHORT 4		COHORT 5		COHORT 6		COHORT 7		COHORT 8		COHORT 9	
MEASURE	GRADE	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT
	K							332	447	253	428	286	339	225	316	X	X	X	X		
	1							344	508	338	442	320	244	X	X	X	X				
	2							396	444	296	329	X	X	X	X						
	3							404	349	X	X	X	X								
	4																				
	5																				
	6																				

X indicates data are keypunched but not in the computer.

(continues)

Table 8 (cont.)

Sample Size by Measure, Cohort, Grade, and Treatment for the Fordham (Atlanta) Data Base

MEASURE	GRADE	COHORT 0		COHORT 1		COHORT 2		COHORT 3		COHORT 4		COHORT 5		COHORT 6		COHORT 7		COHORT 8		COHORT 9	
		FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT
GASC	1											246	91								
	2							268	61	198	83	251	115	207	169						
	3							269	85	212	128	190	151								
	4							229	168	172	127										
	5					283	78	225	112												
MRT	1			229	48	247	91					166	119	187	89						
MAT (SRI)	3							398	127												
Coopersmith	3							218	151	283	213	254	177								
IARS	3							217	153	282	209	253	177								
ITBS	1							304	4[illegible]7	288	367	292	229	*	*	*	*	*	*	*	*
	2							354	342	280	305	*	*	*	*	*	*	*	*		
	3							357	467	*	*	*	*	*	*	*	*				
	4							*	*	*	*	*	*	*	*						
COI	3							90													

(concluded)

*Available from Atlanta P.S. System.

Table 8 (cont.)

Sample Size by Measure, Cohort, Grade, and Treatment for the Fordham (Atlanta) Data Base

		COHORT 0		COHORT 1		COHORT 2		COHORT 3		COHORT 4		COHORT 5		COHORT 6		COHORT 7		COHORT 8		COHORT 9	
MEASURE	GRADE	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT
Basecheck	K-Pre									204	270	244	177	112	189	**	**	**	**	**	**
	Post									191	237	242	146	174	201	**	**	**	**	**	**
BMS	K									210*		285*		231*							
	1							316*		285*		304*									
	2					335*		337*		231*											
	3			434*		389*		333*													
PST	K-Pre							311*		184				152	**	**	**	**	**	**	**
	Post							393*		195	61	233	35	171	**	**	**	**	**	**	**
	1-Pre							285		265		260				**	**	**	**		
	Post							295	58	282	19	273				**	**	**	**		
	2-Pre							318		245						**	**				
	Post							305	25	252				283	81	**	**				
	3-Pre							319													
	Post							263				260	81	**	**						
	4-Pre																				
	Post							60	44	**	**	**	**								
	5-Pre																				
	Post					68	36	**	**	**	**										
TASC	1											247	92					**	**		
	2							269	62	203	84	259	118	207	169						
	3							263	85	209	128	190	151								
	4							233	173	172	127										
	5					290	77	225	112												

*Indicates data that are currently not separated into FT/NFT.

and put in the Fordham system. Fordham feels that the test administration conducted by the Atlanta Public Schools was good, although some mixed-level testing may not have been appropriate. Measures collected in New York as part of the national evaluation are also of good quality and offer a wide base for various analyses.

Form and Organization

The New York data reside at Fordham on computerized data files longitudinally organized by child. This data set has already been used by Fordham in several analyses. It is clean and can be transmitted immediately onto tape.

Part of the Atlanta data exist in uncleaned cross-sectional files, which must be merged with the Fordham-maintained portion of the data set. Matching will not be difficult because the Fordham and Atlanta files use the same child index system. However, substantial effort will be required to produce a clean longitudinal data base from these data.

Availability

Data in Fordham's possession (including all of the New York data and portions of the Atlanta data) are available, subject to the usual controls over release of data and confidentiality. Atlanta's cooperation would be essential to making maximum use of their data. They have not yet indicated the availability of these data, but Atlanta has cooperated with external research efforts in the past. Every large school system is, of course, under an obligation to maintain proprietary rights to data they have collected. Any study using this data base should involve Atlanta personnel as substantive consultants as well as data specialists.

Other Technical Considerations

In Atlanta, children participated in both FT and Title I programs for a period of time. This created some conflict with the Fordham curriculum and may have affected implementation.

5.2.3 Preliminary Study Plan

The Fordham data set affords one of the best opportunities for LE studies. There are large samples over several cohorts, well-matched comparison groups, full demographic data, six years of longitudinal achievement data (see Table 9) with the same measure (different forms), school readiness data, and an array of cognitive and affective measures. The data fit well into the cumulative panel data approach previously presented.

Preparation of Files

We recommend that preparation of files be conducted by Fordham personnel, who are the most familiar with the data, existing files, and existing programs. The following steps should be included:

- The present Atlanta data base maintained by Fordham was cleaned by elimination of problem records rather than by resolution of the problems (e.g., by references to primary materials). By carrying out this quality improvement procedure (already done for the New York file), samples about 3-4 times as large as those on the current file can be obtained. There is a trade-off between time and money spent on this task versus the resultant sample size.
- Data now in the Atlanta data system but not in the Fordham-maintained file must be appended.
- Year files, i.e., child records for particular years not now in the Fordham, Atlanta or New York longitudinal files, must be appended.
- We recommend that the Atlanta file be split randomly into two 50% samples. These can then be used for cross-validation during the model-building process as well as afterward. Sample sizes are large enough to support this approach.

Preliminary Analyses

Determination of the standard of comparison, intergroup adjustments, estimation of the principal sources of variance, psychometric properties of the data, etc., should be performed along the lines suggested in Section 5.1. The outcome should be a final analytic plan for the selection of measures and subsets of the data to be included in analyses, together with warnings of potential problems with the validity of hypothesis tests, etc. It is expected that analyses will use longitudinal samples with the child as the unit of analysis, and Atlanta and New York should be analyzed as separate sites.

Because of the strong panel nature of the data, and the proposed cumulative panel analysis to form the main conclusons, it is important to explore the time series properties of the data in a preliminary way. There are long time series of Iowa Tests of Basic Skills, Phonic Skill Test, absence, and mobility data. Questions that should be addressed are:

Table 9

Measures Available on the Fordham Data Base by Grade and Measure (Atlanta Project, Cohort IV)

	Grade	PS	K	1	2	3	4	5	6
Measure	Year		1972-73	1973-74	1974-75	1975-76	1976-77	1977-78	1978-79
Tests									
Basecheck			XX*						
PST			XX*	XX*	XX*				
TASC/GASC					X	X	X		
Coopersmith						X			
IARS						X			
ITBS				X	X	X	X	X	X
Treatment									
Preschool Exp.		X							
FT Exposure			X	X	X	X			
FT Implementation						X			
Post-FT Treat.							X	X	X
Process									
Mobility			X	X	X	X	X	X	X
Absence			X	X	X	X	X	X	X
Grade Retention			X	X	X	X	X	X	X
Demographic		X	X	X	X	X	X	X	X

---

*Pre and posttests are available within grade.

- What are the recognizable longitudinal patterns in these data and what is their variation across FT/NFT, schools, and other groupings? What is the effect of different metrics (grade equivalents, normal curve equivalents, percentiles, etc.) on these patterns? What do the data suggest as the best metric and the best measures of post-FT effects?

Analyses should examine trends (of rolling averages and first differences as well as of simple values), cross-correlations of series, etc. These analyses are standard in engineering practice but rarely performed in educational analyses, mainly because long time series are much less often seen in educational data sets. These time series questions suggest that descriptive studies be used formatively, not as evaluations of post-FT effects. They will be used to choose metrics for LE variables.

Other questions to be addressed at this stage are:

- What are the differences among cohorts in Atlanta and New York? Are there data that should be discarded because of anomalies and documented threats to validity? Where may cohorts be pooled to increase statistical power? Are there cohort effects, and, if so, should some cohorts be discarded from the sample?

A good deal of descriptive data, sufficient for a useful and detailed report, will be made available by answering these questions. It will set the stage for a more sophisticated panel data approach.

Summative Analyses

The following discussion is ordered according to the types of questions that can be addressed by the data at each of several grade levels.

Overall Questions. The suggested panel approach to the data has been explained previously. The model-building process should be conducted simultaneously on the split halves of the Atlanta file and the entire New York file. Cross-validation enters as an intrinsic part of the process.

Fordham has regarded absence as the cause of poor performance, but it may be that performance and absence have a common cause or that poor performance leads to increased absence. It may be impossible to determine this, but a definite conclusion would be an important contribution to knowledge.

Specific questions that should be addressed include the following:

- Are there any lead-lag relations among time series of data (Phonic Skill Test, Iowa Test of Basic Skills, Test Anxiety Scale, General Anxiety Scale, process data)? Can causal relations among these variables be determined? In particular, what is the causal relation between school absence rates and performance on tests?

- Are there identifiable patterns of effects over time? Specifically, what is the longitudinal pattern of preschool effects? Of FT effects?
- What general conclusions can be drawn about the relative importance of prior and concurrent determinants of outcomes? That is, how important is the current school year compared to the entire previous history of the child?

Kindergarten Questions.

- What are the interrelations among entering-K measures? How do these depend on causally prior preschool treatment and demographic measures? How does FT/NFT treatment depend on these variables?
- How do mobility and absence during kindergarten depend on entering-K measures (achievement, anxiety, phonic skills)?
- How do exiting-K measures (retention, achievement, Phonic Skills Test) depend on entering-K measures (achievement, anxiety, phonic skills)?

The analysis would result in a three-stage linear model, where time (entering-K, during-K, exiting-K) establishes causal priority. The intent here is to learn whether absence and mobility tend to be associated with status measures and whether they operate independently as predictors of exiting-K performance (i.e., beyond their role as status measures). Analysis would begin with multiple regression and correlation models and go on to LISREL causal modeling. Weaknesses and ambiguities in identifying a model should be made clear. These analyses should be informed by substantive theory as much as possible. Models should be constructed not as totally empirical exercises but as expressions of developmental theory whenever feasible.

First through Third Grade Questions.

- What are the concurrent interrelations of the Phonic Skills Test, Test Anxiety Scale, General Anxiety Scale, Iowa Tests of Basic Skills, FT exposure, mobility, and absence? Are there identifiable directions of causality among these variables? How do they depend on previous data panels?
- What can be concluded about the cumulative build-up (if any) of FT effects in grades 1-3? Do FT dropouts appear to be adversely affected in grades 1-3? Do they differ substantially

from non-dropouts? What conclusions can be drawn about the meaningfulness of total FT exposure as an independent variable?

- What is the nature of the one-time-only Coopersmith and Intellectual Achievement Responsibility Scale measures? Are they more nearly status measures associated with family background, or are they determined by educational treatment? Should they be retained in a panel data model?

Post-FT Questions

- What can be determined about the importance of post-FT environment as a determinant of post-FT performance? Do FT/NFT and post-FT environment interact? The answers have relevance to the policy issue of FT program extension to grades 4-6.
- What is the nature of post-FT academic performance? How does the process depend on K-3 data panels and the child's initial status? In particular, what is the relation to FT exposure, FT implementation, and preschool exposure?
- What is the continuing relation among non-academic measures (absence, mobility, etc.) and academic achievement?

Validation

Models and conclusions for the preceding sets of questions would be based on ongoing cross-validation via split halves of the data (for Atlanta only). Thus, no grossly invalid results should survive to this stage. If resources permit, we recommend a full-scale analysis of sensitivity and robustness of conclusions. Any abnormally high influence of small parts of the data should be determined. Specific techniques that can be used include linear regression diagnostics (Belsley, Kuh and Welsch, forthcoming), robust regression techniques and the jackknife (Mosteller and Tukey 1977). Use of such methods would greatly enhance credibility and acceptability of conclusions.

Generalizability

The generalizability of findings to broader universes (sites, other FT approaches) cannot be assessed statistically from the data available in the Fordham data base. However, one can compare results here with those from other sponsors (including sponsor studies) to develop informed opinions of generalizability. This should be addressed in the study's final report.

Organization and Management

USOE could select a prime contractor who would subcontract with Fordham and Atlanta for data preparation. This would be the surest way of completing the work within about one and a half years. Fordham personnel have indicated their preference for an arrangement of this type.

The preparation of a clean Atlanta file at Fordham will take about three months. The New York file could be analyzed first in an exploratory way to decide on final research strategies. The resulting plan would then be used for both the New York and Atlanta data when the latter became available. Estimates of required resources are provided in Table 10.

5.3 University of Oregon Data Base

5.3.1 Educational Model and Later Effects Hypotheses

The University of Oregon employs the Direct Instruction model and the DISTAR curriculum. Criterion-referenced test data (taken every two weeks in each of three areas) are used to monitor progress. The central purpose is to ensure that achievement gains for disadvantaged children are at national norms. There are no important "intervening variables." Oregon believes that positive affect will follow as a byproduct of its methods, but this is not a direct goal. Oregon teaches some systematic problem-solving strategies, especially in decoding, math problem solving, and language. It is hypothesized that these strategies would not be lost in later grades.

Some fade-out in post-FT grades (decreased learning rate) is to be expected for children in the Oregon FT model. Oregon researchers feel that standard curricula in grades 4-6 do not build well on FT K-3 strategies or add new ones. They expect post-FT learning to occur at rates slower than the national norm but faster than that of similarly disadvantaged children. The basic skills learned in FT will continue to be of some value as a learning base. Oregon does not claim to produce major changes in domains other than achievement, but to some extent expects enhanced self-image (following success) and, consequently, some continuing effect on learning rate. These are hypothesized to follow as correlates of increased competence in school.

5.3.2 General Description of the Data Base

Oregon maintains a centralized longitudinal data file for all chilren in its projects. In January 1979 there were a total of 13,849 children who had finished FT before 1976 for whom at least two testings (K, 1, 2 or 3) were available, thus furnishing an opportunity for gains analyses. For approximately 2,500 children, entering (pretest) and exiting (post grade 3) scores

82

Table 10

Estimated Resources Required For the Proposed Fordham Study

Prime Contract		
	Direct Costs	
	Computer	$ 25,000
	Other (telephone, Xerox, printing, graphics, consultants, travel)	10,000
	Total	$ 35,000
	Labor	
	Senior Analysts	24 months
	Junior Analysts	12 months
	Support Staff	12 months
	Total	48 months
Subcontract		
	Data Preparation	
	Atlanta Public Schools	$ 10,000
	Fordham University	40,000
	Total	$ 50,000

are available. The large diminution of sample size--13,849 to 2,500-- is due to low testing rates for entering children at the beginning of the national evaluation.

Measures Available

Measures for grades K-3 include the Wide Range Achievement Test, the Slosson Intelligence Test, and, beginning in spring 1972, the Metropolitan Achievement Test. Testing times were fall (entering) and spring thereafter. Attendance data, demographic data, and Head Start participation data are also available. Table 11 exhibits the temporal patterns of data available for three cohorts where post-FT analyses are possible. Table 12, adapted from Becker and Engelmann (1978), lists all measures, while Table 13 presents Oregon's assessment of each measure's relevance to the Direct Instruction model. There are no pretests for the Metropolitan Achievement Test; instead, the Wide Range Achievement Test and the Slosson Intelligence Test must be used as statistical controls for initial status. Oregon does not feel that the absence of noncognitive measures is a weakness. The measures obtained by Oregon are appropriate for their model, and they distrust the reliability and usefulness of such tests as the Coopersmith, the Intellectual Achievement Responsibility Scale, and the Ravens. In spring 1975 and again in spring 1976, Oregon collected scores on the Wide Range Achievement Test (Levels 1 and 2) and the Metropolitan Achievement Test (Intermediate) for fifth and sixth grade children in sites where control groups could be found and new testing was feasible. LE analyses would be based on these data.

Comparison Groups

Comparison groups are available in seven sites. FT and NFT groups are reasonably well matched, with some exceptions, as exhibited in Table 14. Smithville and Flint are poorly matched with regard to income, Tupelo with regard to race, and E. St.Louis with regard to mother's education. The University of Oregon has customarily made covariable adjustments in its own analyses.

Sample Sizes

In 1975, there were 59 Cohort I-EK and 357 Cohort II-EF children tested in fifth grade and 285 Cohort I-EF in sixth grade. There were 442 control fifth graders and 325 control sixth graders. The elimination of

Table 11

Measures Available on the Oregon Data Base by Cohort and Grade

	Grade						
Cohort	Fall 1	Spring 1	Spring 2	Spring 3	Spring 4	Spring 5	Spring 6
I	1969-70 WRAT SIT	WRAT SIT	WRAT SIT	WRAT SIT MAT	--	--	WRAT (1)* WRAT (2) MAT
II	1970-71 WRAT SIT	WRAT SIT	WRAT SIT MAT	WRAT MAT	--	WRAT (1) WRAT (2) MAT	-- WRAT (2)** MAT
III	1971-72 WRAT SIT	WRAT SIT	WRAT MAT	WRAT SIT MAT	--	WRAT (1)** MAT	--
All	Absence Demographic	Retention Mobility Absence Demographic	Retention Mobility Absence Demographic	Retention Mobility Absence Demographic	--	Demographic	Demographic

---

*Spelling section omitted on all post-FT WRAT tests.

**Math section omitted in 1976 post-FT WRAT tests.

Table 12

Measures Available on the Oregon Data Base*

1. Wide Range Achievement Test
    - (a) Reading (decoding words)
    - (b) Arithmetic
    - (c) Spelling
2. Metropolitan Achievement Test
    - (a) Primary I
        - (1) Word Knowledge
        - (2) Word Analysis
        - (3) Reading
        - (4) Total Reading
        - (5) Total Math
    - (b) Primary II
        - (1) Word Knowledge
        - (2) Word Analysis
        - (3) Reading
        - (4) Total Reading
        - (5) Spelling
        - (6) Language
        - (7) Math Computation
        - (8) Math Concepts
        - (9) Math Problem Solving
        - (10) Total Math
    - (c) Elementary (Same scores as Primary II less Word Analysis)
    - (d) Intermediate (Same as Elementary plus Science)
3. Slosson Intelligence Test
4. Attendance
5. Lessons taught in programs (Reading, Arithmetic, and Language)
6. Criterion-referenced tests (Reading, Arithmetic, and Language) for the years 1970 to 1973
7. Teacher-aide background form
8. Student background form
9. Individual site parent and teacher questionnaires
10. Teacher and aide evaluations by supervision
11. Student performance measures in experimental studies of instruction

---

*From Becker and Engelmann (1978).

Table 13

Relation of Oregon Objectives to Measures on Oregon Data Base*

Objective	Measures	In National Evaluation?	In Site Evaluations?	Appropriateness of Measures (5=very appropriate 1=barely suitable)
1. Norm-Referenced Gains:				
(a) Reading	WRAT-Reading	Yes	Yes	5 (decoding)
	MAT-Reading			4-levels 1, 2 3-level 3 (Comprehension)
(b) Math	WRAT-Math	Yes	Yes	4-levels 1, 3 2-level 2
	MAT-Math	Yes	Yes	4-levels 1, 2, 3
(c) Supplemental Measures	WRAT-Spelling	No	No	3
	MAT-Spelling	Yes	Yes	3
	MAT-Language	Yes	Yes	5
2. IQ Gains:	SIT	No	Yes	4
3. Instruction of Low IQ Students:				
(a) Reading	WRAT-Reading	Yes	Yes	5
	MAT-Reading	Yes	Yes	3
(b) Math	WRAT-Math	Yes	Yes	4-levels 1, 3 2-level 2
	MAT-Math	Yes	Yes	4
4. Fifth and Sixth Grade Study:				
(a) Reading	WRAT-Reading			5 (decoding)
	MAT-Reading			4 (comprehension)
(b) Math	WRAT-Math			4
	MAT-Math			4
(c) Supplemental Measures	MAT-Science			4
	MAT-Language			3
	MAT-Spelling			3
5. Effects of Head Start (Planned Variation):				
(a) Reading	WRAT-Reading			5 (decoding)
	MAT-Reading			4 (comprehension)

(continues)

---

*From Becker and Engelmann (1978).

Table 13
(concluded)

Objective	Measures	In National Evaluation?	In Site Evaluations?	Appropriateness of Measures
(b) Math	WRAT-Math			4
(c) Supplemental Measures	MAT-Math			4
	SIT-IQ gain			4
	MAT-Spelling			3
6. Relations of Process and Outcome:				
Experimental Studies	Student Rate and Accuracy	No	No	?
	Time on Task	No	No	?
	Amount Mastered	No	No	?

Table 14

Comparison of FT and NFT Groups on Selected Fifth Grade Demographic Variables by Site for the Oregon Data Base

Site	Mother's Ed. Scale FT	Mother's Ed. Scale NFT	Percentage Low-Income FT	Percentage Low-Income NFT	Percentage Non-White FT	Percentage Non-White NFT
Flippin, Ak.	4.62	4.45	.33	.35	.00	.00
E. St.Louis, Il.	4.35	3.68	.93	.98	1.00	1.00
Smithville, Tenn.	3.30	3.70	.87	.54	.17	.00
Uvalde, Tx.	2.86	2.34	.87	.95	1.00	.95
Dayton, Oh.	4.75	5.00	.85	.71	1.00	.91
Tupelo, Ms.	3.22	3.50	.95	.97	.93	.53
Flint, Mi.	--	--	.94	.43	.94	.77

nearly all EK children occurred because most of the sites where cooperation could be secured and comparison groups selected were located in the South and Midwest, where kindergartens were uncommon in 1969-71. Sample sizes by site and treatment group are given in Table 15. Each child in this sample was present both in the spring of grade 3 and at the post-FT grade level indicated.

Demographic Data

Demographic data include ethnic group, home language, date of birth, sex, poverty code, and Head Start experience.

Test Administration

Test data are excellent in terms of quality control over test administration and well-organized file construction.

Form and Organization

The K-3 data file is a fully centralized longitudinal computerized system. In 1975, there were seven sites and about 700 post-FT children who could be traced back to existing longitudinal files. In 1976, there were only six sites and somewhat fewer children. There is presumably a substantial overlap between these testings (5th in 1975, 6th in 1976) so that a linked file could be constructed, but such a file does not now exist.

Availability

The data are available for reanalysis, subject to usual controls for children's privacy.

Other Technical Considerations

None.

5.3.3 Preliminary Study Plan

In USOE's national evaluation, Oregon appeared to be the most successful sponsor in obtaining improved performance on achievement tests. This characteristic, in combination with Oregon's reputation as a "basic skills" sponsor, makes their data particularly interesting for a study of fade-out and persistence of effects. The data provide a strong base for a study of reading and math performance through grade 6. Such studies would be highly

Table 15

Longitudinal Sample Size by Site and Grade for the Oregon Data Base

	Spring 1975				Spring 1976			
	FT		NFT		FT		NFT	
Site	5th	6th	5th	6th	5th	6th	5th	6th
Flippin, Ak.	24*	18*	57	61	4*	6*	13	16
E. St.Louis, Ill.	43	47	45	44	45	37	44	48
Smithville, Tenn.	47	48	51	47	71	38	38	46
Uvalde, Tx.	117	101	86	101	103	108	74	73
Dayton, Oh.	104	33	87	47	--	--	--	--
Tupelo, Ms.	46	38	34	25	56	55	42	33
Flint, Mi.	35*	--	82	--	46*	22*	43	24
Total	416	285	442	325	325	226	259	240

*Started FT in kindergarten. All others began in first grade.

relevant to the sponsor's contentions about the inferiority of the standard curriculum in grades 4-6 and are of policy interest with regard to possible later grades extensions of FT. Oregon has conducted its own quasi-longitudinal analyses, which indicate that positive FT effects are still significant, although diminished, in fifth and sixth grade, i.e., that their hypotheses about post-FT effects appear justified. These analyses are of high quality and provide important contributions to knowledge about FT.

The purposes of a reanalysis of the data would be to provide judgment by uninvolved analysts and to extend the Oregon analyses to include analysis of absence, retention, and mobility data; to perform multivariate longitudinal analyses that capitalize on the time series nature of the data and their evolution in time, i.e., a cumulative panel data analysis; and to conduct a study of site effects and an analysis of the sensitivity and robustness of effects.

Because of the simplicity of the data set, lack of large data reduction costs, and usefulness of existing sponsor analyses, the Oregon study would be of relatively modest size and cost.

Preparation of Files

Since sample sizes are small in the only two entering-K sites with LE data (Flippin, Flint), the analysis should concentrate on the five entering-first sites, where sample sizes range from about 100 to 200 per site (over two cohorts). The measures listed in Table 12 may be easily pulled from Oregon's files. It is possible to extend the range of measures by attaching mobility data and absence/retention data from public school records for grades 4-6. This would require manual effort at each school, but would be relatively inexpensive for the modest sample size. A child-by-child longitudinal file would be the product of this effort.

Preliminary Analyses

In order to formulate summative analytic strategies, some preliminary analytical questions should be asked. Section 5.1 presents general discussions of the origins of variance in outcomes, the unit and level of analysis, standards of comparison, intergroup adjustment strategy, psychometric investigations, longitudinal vs. cross-sectional analysis, etc. Although each of these areas must be reconsidered in formulating a final analytic plan, it is

92

expected that the unit of analysis will be the child, the level of analysis will be the sponsor, the adjustment strategy will be ANCOVA, and analyses will be longitudinal. The following paragraphs offer some further specifics.

Unit and Level of Analysis. A final decision must be made between doing five to seven separate site-level analyses or combining data across sites for enhanced statistical power. Variance components analysis can be used to help decide whether site effects are small enough to allow a simple adjustment strategy to be used to equalize sites. In no case should entering-first and entering-K sites be combined. The unit of analysis must be the child, for the reasons previously cited.

Standard of Comparison. Oregon used ANCOVA techniques, and we recommend the same (specifically, a univariate composite covariable approach); the drawbacks of the method are well known. If the sponsor is selected as the level of analysis, one would have to decide between separate within-site covariable models or a global one if the sponsor is selected as level of analysis. The answer to this would depend on heterogeneity of models from site to site and the usefulness of a simple adjustment for site effect. Answers can be obtained by straightforward hypothesis testing, using the Chow (1960) test.

Longitudinal/Cross-Sectional Analyses. The sample sizes in Table 15 are longitudinal across two points of time (exit from grade 3 and the post-FT test time). If more time points are used (e.g., entering-K pretests), within-site samples may sink to levels that will not support meaningful hypothesis testing. If it is possible to pool across sites, then good longitudinal analyses can probably be supported.

The existing quasi-longitudinal Oregon analyses are very good as far as they go. If technical circumstances do not allow use of longitudinal methods, then it is questionable whether an external reanalysis would contribute much new knowledge. If longitudinal analyses are possible, however, then the entire K-6 process can be studied as a dynamic whole, and new knowledge is likely to be obtained. The answer to the longitudinal/cross-sectional question is thus critical to an expectation of the usefulness of results.

Scales and Subscales. Tables 12 and 13 summarize the subscales available and Oregon's assessment of their relevance to model objectives. In a model as curriculum-oriented and as specifically targeted as Oregon's, these subscales assume major significance. One is particularly interested in transfer from more mechanical areas, such as decoding and computation, to the more conceptual ones, like IQ, math problem solving, etc. There is also the hypothesis that post-FT effects are different in different domains of measurement. Studies by both Oregon and Kansas find, for instance, that math effects are harder to sustain than reading effects. It would be valuable if these differences could be related to the Oregon curriculum and the standard school curriculum. This is, of course, a special interest at Oregon, where it is believed that standard educational practices are inefficient.

In order to provide a base for such analyses, it is important that psychometric analyses of the type discussed in Section 5.1 be conducted.

Summative Analyses

The overall objective is to understand and describe the interrelation of the available measures, their temporal behavior through grade 6, and the determinants of observed outcomes. This objective is similar to that of the proposed Fordham study, but the data will not support a full-scale cumulative panel data approach as suggested for Fordham. Several technical problems interfere. First, an unbroken time series of measures (or even a pretest for the Metropolitan Achievement Test) does not exist; no measures are available for grade 4. Second, sample sizes would probably be reduced to unacceptable levels in grades K-6 for longitudinal analyses that required more than two or three linked measures. However, the Fordham plan represents one kind of ideal toward which an Oregon study should strive.

The most important reason to try to formulate causal and dynamic relations among variables from grades K-3 is to learn something of their meaning as independent variables in LE analyses. There are so many K-3 measures in the data base and so much risk of multicollinearity that there needs to be a model whereby measures are selected or rejected for substantively meaningful reasons.

The candidate independent variables include the following: measures of child's pre-FT status including the Wide Range Achievement Test, the Slosson Intelligence Test, and demographic data; the record of school experiences, including participation in and amount of exposure to Head Start and FT/NFT; and short time series of data from grades K-3, including the Wide

Range Achievement Test, the Slosson Intelligence Test, the Metropolitan Achievement Test, absence, and grade retention. These are measures of success in grades K-3. The dependent measures consist of scales from the Wide Range Achievement Test and the Metropolitan Achievement Test, together with absence and retention data, administered in 5th and 6th grades.

The research questions include the following. The first few are attempts to construct a meaningful set of independent variables for LE analyses.

- What are the relations of spring grade 1 measures to fall grade 1 measures, to pretests, to FT/Head Start program exposure?
- What are the relations of grade 2 measures to spring grade 1? To fall grade 1, controlling for spring grade 1? Is the effect of preschool treatment and fall grade 1 measures attenuated? How severe are multicollinearity problems between spring grade 1 and fall grade 1 domains?
- What overall step-by-step model can be constructed in the manner suggested by the above two research questions? What is implied about the continued importance of earlier variables as they are layered over by more recent educational experiences?
- When sets of subscales are considered as panels, what lead and lag relations appear to emerge? Are there any lead-lag relations between the Wide Range Achievement Test and the Slosson Intelligence Test?

Since the Wide Range Achievement Test is a rather mechanical instrument (highly relevant to Oregon methods), and the Slosson Intelligence Test is conceptual, this question is of substantial interest.* For some educational psychologists, IQ tests are considered especially useful for indicating differences that children bring with them from home to school. Therefore, IQ-adjusted residuals may be of considerable interest. However, in the event that no lead-lag relation emerges, one should not necessarily conclude that none exists, because of marginal statistical power and the possible inappropriateness of the one-year frequency of measurement. The answer is useful only if a lead-lag relation is found. Then weak inferences can be made about direction of causality.

- Are there causal inferences that can be drawn about absence and mobility and their relation to test performance?

By exploring these areas it will be possible to learn something about the dynamic process by which new variance sources replace old ones, and about the nature of remaining dependence on the child's initial status. It will then be possible to address the following questions.

---

*Oregon analyses show possible effects on the Slosson Intelligence Test--despite program concentration on more mechanical skills.

- What primary independent variables should be used in an analysis of post-FT effects of the Oregon model? Ideally? Considering effects on longitudinal sample size? What are their places in a (tentative) causal model of FT/NFT effects? What is the usefulness of length of exposure and absence data?

At this point, one can begin to explore relations of grade 5 and 6 measures to grades K-3 measures (including covariables). Techniques will nearly all be variations of the familiar ANCOVA model. The adjustment problem (between nonequivalent control groups) must be carefully considered. Then the following questions can be answered.

- What FT/NFT effects can be found across the panel of grade 5 or grade 6 measures? What are other major contributors to variance? What rival hypotheses exist for FT/NFT effects--i.e., are there plausible alternative explanations of effects?
- For the subsample of children with both 5th and 6th grade measures, is there a pattern of continued attenuation of effects?
- Are there substantively interpretable patterns of effects? For example, are there explanations within Oregon's model that account for differences across different measures?

Validation

Sensitivity and robustness analyses should be made. Dependence of conclusions on particular parts of the data should be determined (see the Fordham study discussion).

Generalizability

As is the case with the other studies discussed in this report, generalizability of results will depend on substantive arguments and integration of results with other studies rather than on statistical tests.

Organization and Management

This study could be performed in a period of one year, with about two months reserved for data preparation by Oregon. USOE could select a prime analytic contractor who would negotiate a subcontract with Oregon for data preparation. Estimated resources for an Oregon study are displayed in Table 16.

Table 16

Estimated Resource Requirements for the Proposed Oregon Study

Prime Contract	
Direct Costs	
Computer	$12,000
Other (telephone, Xerox, printing, graphics, consultants, travel)	6,000
Total	$18,000
Labor	
Senior Analyst	12 months
Junior Analyst	6 months
Support Staff	6 months
Total	24 months
Subcontract	
Data Preparation (Oregon)*	
Total	$16,000

---

*This is an approximate estimate by Orgeon, including computer, typing, programming, reports, labor, and overhead.

5.4 University of Arizona (Wichita, KS) Data Base

5.4.1 Educational Model and Later Effects Hypotheses

The Tucson Early Education Model has four direct goal areas, including development of a language base, an intellectual base (cognitive conceptual), a motivational base (positive attitudes), and societal arts and skills. The latter area includes reading, writing, math, social skills, and understanding of the democratic process. Implementation of these four goal areas is intended to develop independence, enhance positive self-concept, and provide opportunity to practice decision making. It is hypothesized that effects should emerge in post-FT years as a result of the FT programs. Wichita's own analyses show no striking effects either for grades 4-6 or for grades K-3. If effects are to be found, they will emerge in particular groups of children or by controlling for initial status.

5.4.2 General Description of the Data Base

In only one of the Arizona sites do extensive LE data exist--the Wichita Public Schools. While the Wichita data set is relatively small (index card file), it possesses many of the features of classical designs. Treatment and control groups were selected randomly from a common pool of eligible children and were followed through grade 6. About 100 FT children and 100 NFT children were selected each year.

Measures Available

Measures by cohort for the Wichita data base are given in Table 17. A range of measures exists, and a variety of studies can therefore be conducted. The California Test of Mental Maturity and the Illinois Test of Psycholinguistic Abilities bear directly on Arizona's direct project goals. Arizona also obtained some data on sponsor-constructed measures, including the Productive Language Achievement Test.

Comparison Groups

Several groups of children were followed. A list of Head Start-eligible children was drawn up, based on social worker assessments of educational need. The "Head Start/FT" group was selected from a randomly ordered list of Head Start-eligible children, working from the top down until 100 children were selected. The "Head Start/NFT" group consists of the remainder

Table 17

Test Data on the Wichita Data Base

Test	Cohort	I							II								III			
	Grade	K	1	2			3	4,5,6	K	1	2			3		4,5,6	K	1	2	3,4,5,6
	Test Point	F*	F	F	W*	S*	F	S	F	F	F	W	S	F	S	S	F	F	F	S
CTMM					X							X								
ITBS							X	X						X		X				X
ITPA		X				X			X				X				X			
MAT				X							X								X	
MRT			X							X								X		
SIT									X	X					X		X			

*F=Fall, W=Winter, S=Spring

of the list. The latter group includes children who rejected FT classes, a minor departure from full randomization when rejection rates are low. The Head Start/NFT group had a larger proportion of whites until 1977-78, when the proportion of FT refusers jumped substantially. This occurred because a policy was implemented whereby all FT children were to be bused out of their neighborhoods. The "FT classmates" group consists of non-eligible classmates of Head Start/FT children who received the FT classroom treatment but not the comprehensive (medical, social, and psychological) services. The "low-income" group consists of children on the waiting list for Head Start, but who are less needy than Head Start children. The "non-low-income" group are randomly selected non-low-income children in four paired non-low-income schools.

Sample Sizes

Sample sizes by group are listed in Table 18. These children are fully longitudinal, i.e., were present from entering-K to the grade indicated. The total sample through grade 6, summing across Cohorts I-V at the end of the 1979-80 school year, is projected to be about 175 (Head Start/FT), with roughly comparable numbers in the other groups.

Demographic Data

Standard data such as sex, race, and age are available.

Test Administration and Validity

Except for the Productive Language Assessment Test, data were obtained as part of the city-wide testing program. There are no particular concerns about faulty test administration.

Form and Organization

The data are in a longitudinal card-file data base (hand entries). The file could be coded and punched at little expense.

Availability

The data are available provided that children's names do not leave the Wichita offices.

Other Technical Considerations

None.

Table 18

Sample Size by Grade, Cohort and Treatment Group for the Wichita Data Base

		Treatment Group				
Cohort	Grade	Head Start/ FT	Head Start/ NFT	Low-Income	Non-Low-Income	FT Classmates
I	6	45	31	35	23	26
II	6	35	21	32	20	29
III	5	33	27	25	29	27
IV	4	28	29	--	31	31
V	3	49	91	--	56	60
VI	2	57	66	--	89	80

5.4.3 Preliminary Study Plan

From a methodological point of view, the Wichita data set is probably the best one available across all sponsors. The Wichita Division of Research, Planning and Development Services has conducted extensive analyses, and the expectation for finding positive effects in reanalyses is somewhat dampened by the fact that existing Wichita analyses show no striking effects either for grades K-3 or for grades 4-6. While this may diminish enthusiasm for reanalysis, Arizona hypothesizes that elementary pupils will develop competence in the basic skills, and the reasonably large sample size and strength of the data set make failure to reject the null hypothesis more interesting than usual.

Preparation of Files

The file would be prepared by clerical personnel sent to Wichita, where they would copy existing records onto coding sheets. These sheets would, of course, omit any confidential information. Some cross-checking of records would be required to ensure the highest quality possible. Coding sheets would then be punched onto cards. Files would consist of longitudinal child records organized by cohort.

Preliminary Analyses

Preliminary investigations along the general lines indicated in Section 5.1 should be conducted. The following comments are specific to a Wichita study.

Standard of Comparison. We recommend retaining Wichita's comparison groups (Head Start/NFT, FT classmates, low-income children, non-low-income children) and performing analyses both with and without covariables. Use of covariables for Head Start/FT vs. Head Start/NFT comparisons may correct for unhappy randomization and for a possible small "refusal" effect.

Unit and Level of Analysis. The unit of analysis is the child. The level of analysis should be the cohort for some analyses and cohorts should be pooled for others. Pooled cohort analyses, which would be conducted to boost statistical power, would involve only those cohorts determined to be sufficiently similar.

Psychometric analyses, exploratory data analyses, definition of the analytic data base, etc., are all standard.

Summative Analyses

Wichita analyses cannot follow the cumulative panel data paradigm discussed previously since there is only one measure available for grades 4-6 (the Iowa Tests of Basic Skills). Several interesting measures are available for grades K-3, however, and the analyses should use these variables as controls for later effects. The following specific questions should be asked.

- Entering-K measures include the Illinois Test of Psycholinguistic ability and the Slosson Intelligence Test. What is the relation of the first grade measure (Metropolitan Readiness Test) to these tests? To FT/NFT?

The entering-K measures are conceived of as initial status variables that measure what the child brings to school from home. Analyses would be conventional regressions, and the conclusions would help to evaluate the impact of FT on first grade readiness, as a function of entering status.

- What are the relations of second grade achievement (Metropolitan Achievement Test) and psycholinguistic ability to educational treatment, previous outcomes, and initial status?

Analyses would be similar to those for first grade.

- From grade 3 on, one has a sequence of four measurements using the Iowa Tests of Basic Skills. What are the distinctive patterns in these longitudinal sequences? Can children be classified as decelerators, accelerators, etc.--i.e., how does later success relate to earlier success? Can particular groups of children, defined in terms of initial status or socioeconimic status, be isolated as being particularly affected by treatment?

These are, of course, the most important issues in a LE study. Since Wichita analyses showed few effects, one can only hope to find effects by isolating data subsets or by appropriately controlling for earlier variables. Nearly all analyses would nearly all use multiple regression techniques.

Validation

Validation should be mostly along the lines outlined previously in Section 5.1. However, for Wichita, we would put considerable emphasis on statistical power. Not only are null conclusions likely to be the most common, but power for data subsets may be quite low. If the study "proves" the null hypothesis, then power of hypothesis tests must be carefully considered and quoted in any reports.

Generalizability

Since the data are available only from a single site the generalizability of Wichita analyses is bound to be low.

Organization and Management

This study could be performed in a period of about six months, with about one month reserved for preparation of hard-copy data files by Wichita. As with other studies, USOE could fund a prime analysis contractor that would subcontract with the Wichita Public Schools for data preparation. An estimate of the resources required for a Wichita study are given in Table 19.

5.5 Southwest Educational Development Laboratory Data Base

5.5.1 Educational Model and Later Effects Hypotheses

The SEDL model is a language-oriented approach that concentrates on the use of special curriculum and teacher training materials. There are separate modes for teaching in both English and Spanish (Bilingual Mode) or in English as a second language (ESL Mode). The most important goals for children are to develop oral language and reading skills in Spanish/English (Bilingual Mode) or English (ESL Mode), as well as to develop self-esteem, cognitive skills, recognition and understanding of other cultures, and pride in their own culture. The program seeks to remove the language barrier to learning, while the child develops greater motivation as a result of affective program consequences.

SEDL expects learning to continue at an accelerated rate even after cessation of the FT treatment. The program's main goal is transition to all-English-speaking classrooms. The fraction of time spent in special SEDL curriculum decreases as the child progresses from grade K to grade 3. If the program is successful, a child should continue to achieve at normal rates in the later grades. SEDL does not anticipate sleeper effects.

5.5.2 General Description of the Data Base

Of the seven SEDL sites, six have some LE data varying from site to site. One site, Philadelphia, was described in the earlier report by Goodrich and St.Pierre (March 1979). SEDL was not implemented in Corpus

Table 19

Estimated Resource Requirements for the Proposed Wichita Study

Prime Contract

Direct Costs

Computer	$2,000
Other (telephone, Xerox, printing, graphics, consultants, travel)	3,000
Total	$5,000

Labor

Senior Analyst	6 months
Junior Analyst	3 months
Support Staff	3 months
Total	12 months

Subcontract

Data Preparation (Wichita)

Total	$2,000

Christi until 1974-75, so post-FT measures are not yet available for that site. Data have been collected by the sites and in three cases (San Diego, Cutler-Orosi, and St. Martin Parish) have been forwarded to SEDL. Data from Tulare are to be sent to SEDL in the near future. The Los Angeles data still reside at the site and may not be available for analyses.

Measures Available

Measures include the following:

San Diego, TX	Survey of School-Related Attitudes Spanish Reading Test Self Concept and Motivation Inventory
Los Angeles, CA	Comprehensive Test of Basic Skills Spanish Reading Test Self Concept and Motivation Inventory
Cutler-Orosi, CA	Comprehensive Test of Basic Skills
Tulare, CA	Comprehensive Test of Basic Skills
St. Martin Parish, LA	Metropolitan Achievement Test Self Concept and Motivation Inventory

Comparison Groups

In San Diego, the comparison group is from a neighboring community that is economically advantaged. In Los Angeles, comparison children have been drawn from the same general East Los Angeles neighborhood and are well-matched with the FT group. The St. Martin Parish comparison group differs from the FT group on several dimensions, indicating that the comparison group is relatively advantaged. Comparison groups are not available for Cutler-Orosi or Tulare.

Sample Sizes

Starting with Cohort III, data are available for each year, grades K through 6. After 1978-79, similar data will be available for Cohort IV and after 1979-80, for Cohort V. Sample sizes at entry to FT and ethnic composition for Cohort III are shown in Table 20. Attrition might reduce the sample sizes considerably in Cutler-Orosi (a migrant community) and Los Angeles (due to urban migration). The other sites are relatively stable.

Table 20

Sample Size at Entry to FT and Ethnic Composition by Site for the SEDL Data Base

Site	N	Chicano	Anglo	Black
San Diego, TX	107	100%	--	--
Los Angeles, CA	153	97%	1%	1%
Cutler-Orosi, CA	130	80%	20%	--
Tulare, CA	74	55%	15%	30%
St. Martin Parish, LA	169	--	51%	49%

Demographic Data

Family income, community demographics, age, race, ethnicity, Title I eligibility, Spanish/English-speaking background are all available.

Test Administration and Validity

Tests were administered by sites with no special controls by SEDL.

Form and Organization

Data are mostly on cards in different formats from site to site.

Availability

Data are known to be available from all sites except Los Angeles. In Los Angeles, the question of accessibility has not yet been resolved.

Other Technical Considerations

None.

5.5.3 Preliminary Study Plan

The SEDL program is of special interest as one that combines elements of compensatory education and bilingual education. This approach is applicable only to limited English-speaking or non-English-speaking populations, so results from a study would not be highly generalizable. However, in view of the current great interest in evaluation of bilingual education programs, results from a study of the SEDL FT program would be exceptionally interesting and policy-relevant.

The research questions that can be addressed are constrained by the data available and the characteristics of available comparison groups. Ideally, one would want valid scales of English-language proficiency, self-concept, and academic performance in a range of content areas. For Spanish-background children, one would also want a test of Spanish-language proficiency. The problem, then, is one of determining how well SEDL children perform in these areas relative to children of the same culture, language variety, and socioeconomic status who: receive no special treatment, receive monolingual compensatory treatment, or receive different bilingual education treatments. The preliminary investigation would determine which of the three sites with comparison groups and adequate sample sizes (San Diego, Los Angeles, St. Martin Parish) would help answer any of these questions and would specify the questions that can be answered. In the following, we assume that the Los Angeles data will be available, although this is not yet clear.

While the general study design follows that set forth in Section 5.1, the SEDL study stands somewhat apart because of the special need to consider cultural and linguistic issues.

Preliminary Analyses

First, it will be important to examine the validity of academic achievement tests available (e.g., Comprehensive Test of Basic Skills) as measures of performance for Spanish-background children. In order to interpret the scores, one must understand to what degree these tests measure English-language proficiency and to what degree they measure knowledge of content. It is also necessary to examine the validity of the Spanish Reading Test and the Self Concept and Motivation Inventory.

A third task of preliminary analysis is to conduct a detailed examination of the background characteristics of FT and NFT children. These should center on socioeconomic status, entering linguistic proficiency, language varieties, and, most importantly, the educational treatment of NFT children. The purpose of these investigations would be to specify the research questions that could be answered in each of the three potential sites.

The most important independent variables are those that describe entering language ability in English and Spanish, since it is language development that the program stresses most highly. It does not appear that

108

good measures of entering language ability are available. Without such measures, one must assume that the language abilities of comparison and FT groups are the same. The study should use local comparison groups matched for ethnicity and family income, since these are probably the best available determinants of language ability. The two sites where local comparison groups now appear adequate are Los Angeles and San Diego, but St. Martin Parish should also be examined.

A further task of preliminary analysis is to determine the specific research questions that can be answered. In which sites? Which sites should be included in the analyses?

SEDL indicates that matching of FT and NFT groups in terms of socioeconomic status is good in Los Angeles, only fair in San Diego, and perhaps poor in St. Martin Parish. It may be possible to isolate better matching subsets of NFT groups, however. The educational treatments that NFT children receive are not known. The knowledge base acquired through the preliminary analyses will indicate what research questions can be addressed and in which sites.

Because educational needs of Mexican, Anglo, and Black children differ substantially, all analyses should be conducted within ethnic group. San Diego, and Los Angeles are such different environments that these data should not be pooled either. Thus the study consists of the following potential units:

Unit	N
San Diego/Chicano	107
Los Angeles/Chicano	150
St. Martin Parish/Anglo	86
St. Martin Parish/Black	83

These sample sizes are distinctly marginal, and the child is the only feasible unit of analysis. Data from later cohorts could possibly be used to bolster sample sizes. Cohort pooling is, of course, problematic because of changing class composition and changes in the program. Since each study will be rooted in an ethnographic, sociolinguistic context, the results would be integrated through verbal reasoning processes.

The discussion of longitudinal/quasi-longitudinal designs, routine psychometric analyses, exploratory data analysis, definition of the analytic data base, and generalizability that were presented in Section 5.1 all apply here.

Summative Analysis

The basic approach to summative analysis set forth in Section 5.1 can be used for the SEDL data, since sequences of similar measures are available for grades K-6. Questions of particular interest would involve the temporal relations among measures that included:

(1) English-language proficiency;
(2) Spanish-language proficiency (Los Angeles, San Diego);
(3) Academic performance in content areas; and
(4) Self-concept.

It may be difficult to construct credible and valid measures of English proficiency and academic performance from the achievement tests available. If this cannot be done, interest in the analysis would be considerably lessened because the analysis will not address the central goal of the SEDL program. If it can be done, then very interesting cumulative panel analyses are possible. Note, however, that sample sizes are small and, hence, lead-lag relations will be difficult to detect. Therefore, it is possible that little of interest will emerge. If cumulative panel analyses do not yield results of interest, then the data will be analyzed through conventional quasi-experimental techniques, treating data from grades 4-6 separately in parallel analyses.

Generalizability

Generalizability of the study results will be restricted because of the special population served by SEDL, the special nature of their program, and the limited number of sites available for study (potentially only a single site).

Organization and Management

The amount of time required to perform this study will depend on the speed with which the data can be prepared by the sites. This might involve a complicated process, and we suggest allowing about three months for

Table 21

Estimated Resource Requirements for the Proposed SEDL Study

Prime Contract

Direct Costs

Computer	$3,000
Other (telephone, Xerox, printing, graphics, consultants, travel)	3,000
Total	$6,000

Labor

Senior Analyst	6 months
Junior Analyst	3 months
Support Staff	3 months
Total	12 months

Subcontract

Data Preparation (SEDL)

Total	$4,000

data preparation and six months for analysis and reporting. A prime contractor should subcontract either with SEDL or directly with the involved school districts for data preparation. An estimate of required resources for a SEDL study are given in Table 21.

5.6 Georgia State University Data Base

5.6.1 Educational Model and Later Effects Hypotheses

Georgia State's Parent Supported Diagnostic Model has two components: home and classroom. Through home visits and teacher training, the sponsor tries to improve the quality of adult-child interactions at home and at school. Improved attitudes (toward self, home, and school) and learning to learn are emphasized. Intensity of involvement and time on task are thought to be the major inputs to enhanced achievement; i.e., the model is process-oriented. Enhanced achievement is posited to follow from enhanced quality of the interaction process. Quality and amount of contact time are stressed. Both achievement and interaction are direct goals of the model.

It is hypothesized that the quality of the classroom environment will deteriorate in grade 4, while, at home, the changed environment will persist. Some fade-out of effects can therefore be expected due to deterioration of the school environment; but, since home environmental effects are hypothesized to persist, FT achievement may remain superior to NFT in later grades. Delayed or "sleeper" effects, which might be expected from a home-based model, are not expected by the Georgia State staff. Georgia State researchers have found evidence of a fourth grade trough followed by a fifth grade rebound.

5.6.2 General Description of the Data Base

Georgia State possesses a fully computerized file of California Achievement Test data collected in a single site from grade 1 through grade 6, organized both longitudinally and cross-sectionally. All data are from Natchitoches Parish, the only site where the model was fully implemented. No extensive data exist that are relevant to the immediate desired outcomes of the model, i.e., classroom and home interaction processes. The Georgia State data files can be extended by annexing additional data available from the Natchitoches Parish Public School system.

112

Measures Available

Georgia State's achievement measure is the California Achievement Test, in grade-appropriate forms. The data base also contains attendance data and some home and school interaction process data taken in systematically scheduled observations. Criterion-referenced test data are also available.

Comparison Groups

The FT data came from six geographically separated schools in six wards of the Parish. There are three NFT schools in three other wards. Most of these schools are *de facto* segragated, and the FT eligibility matches are only fair. They appear to be useful comparison schools, however, provided that bad mismatches are discarded and (possibly) some adjustment strategy is employed. There is also a grades 4-5-6 "advance control group" that could be used in descriptive studies.

Sample Sizes

Data from grades 4-6 are available at Georgia State University for Cohorts I-III, with cross-sectional sample sizes given in Table 22. Data from grades 1-3 are available for most subjects, and (as discussed above) the data set may be augmented by data now available only at the site. The small sample sizes for Cohort I, grade 4 are due to probability sampling during 1972-1973. The data may be viewed as sequences (i.e., time series) of measures on longitudinal growth curves. The total number of trajectories extending to grade 5 or 6 (allowing for attrition, but accepting some missing values) is estimated to be on the order of 200, sufficient for analyses if the data are pooled across cohorts. Some random sampling was involved in constructing the data base.

Table 22

Sample Size by Grade, Treatment, and Cohort for the Georgia State Data Base

		Cohort		
Grade	Treatment	I	II	III
4	FT	45	106	127
	NFT	16	59	53
5	FT	154	109	149
	NFT	58	49	59
6	FT	151	---	---
	NFT	51	---	---

Demographic Data

Data available include date of birth, sex, race, education, preschool experience prior to FT, number of siblings, and birth order.

Test Administration and Validity

Testing was conducted by teachers, but paid for and monitored on site by Georgia State University. Administration and scoring was subject to the usual problems of on-site inefficiency.

Form and Organization

The data are organized into two files. One is by cell (e.g., by cohort and grade), suitable for cross-sectional analysis. The other is longitudinal, by child. The data set, extended by new data from the site, can be prepared for external use by Georgia State researchers.

Availability

The data are in a computer-readable data base and are available.

Other Technical Considerations

There is little diffusion of the Georgia State model within school. In fact, the FT/NFT classroom contrast is strong and may lead to transitional problems. Attrition is not severe; about 80% of the initial sample is available longitudinally.

5.6.3 Preliminary Study Plan

The Georgia State model is implemented fully in only one site, and for this reason it was not included in the national evaluation. A major problem is that the site (Natchitoches Parish, LA) is so highly specialized that results of studies would have little generalizability. A study of the Georgia State data must, therefore, be regarded essentially as case studies, although quantitative methods will be used.

Preparation of Files

The existing computer files of longitudinal California Achievement Test data are already available but include no measures of impact on home environment or classroom process. Thus, one cannot address questions about the impact of the program on process at home or school, or the relation of inter action process and achievement. To remedy this, available non-computerized interaction process data, attendance data, and some criterion-referenced data should be attached to the existing files. If this is done, process/product analyses bearing directly on model hypotheses could be conducted. Since interest in the findings would be enhanced considerably by such analyses, this data preparation work should be conducted if a study is to be performed at all.

Preliminary Analyses

The basic plan set forth in Section 5.1 should be followed. Some specific issues concerning the Natchitoches Parish data are as follows.

With only six FT and three NFT schools, the child is the only reasonable choice as the unit of analysis. Selection of a level of analysis is not so clear, and depends on closer examination of child characteristics in the six FT and three NFT schools. Most of these schools are *de facto* segregated and FT eligibility matches are only fair. We recommend against pooling of analyses based on largely black and largely white schools, since these schools are likely to differ considerably as educational environments. At the same time, it is not possible to match each of the six FT schools with a different NFT school. We see two reasonable alternatives. First, analyses could be conducted separately for children in black, white, and integrated schools. In each case, at least one comparison school, not badly matched for socioeconomic status, must be available. If one is not available in a particular racial category, then that analysis must be discarded. Analyses would use ANCOVA to help correct for child-level socioeconomic differences. Second, analyses could be conducted separately for black and white children. In each case (black and white), a comparison group would be selected from the set of three NFT schools. By selection, socioeconomic matching can be secured. Analyses would be adjusted by standard ANCOVA methods.

Selection of a design plan must be based in part on a detailed examination of the characteristics of the nine schools and the children in them. It would be valuable to conduct anthropological and/or ethnographic investigations to assist in choosing a design and understanding the results to be obtained.

Summative Analyses

Three forms of analyses could be conducted:

- The standard ANCOVA approach, applied to California Achievement Test scores from grades 4-6. Analyses could take a variety of approaches to control for entering status and exiting (grade 3) academic performance.
- The cumulative panel data approach can be taken, modified for the univariable sequence of California Achievement Test scores. Of course, no causal hypotheses can be tested. Instead, one could use a systematic time series approach to the assessment of cumulation, persistence, and/or erosion of effects. Absence data could also be introduced in these analyses.
- Process/product analyses could determine relations between interaction variables to be constructed and California Achievement Test measures of performance.

Interaction variables would indicate, crudely, "amount" of treatment. No control groups would be required in these analyses. An ANCOVA regression approach would be taken.

The results of these analyses could be presented, together with a narrative explanation of social and educational conditions in Natchitoches Parish. As remarked earlier, this is a quantitative case study, and the quantitative conclusions should be bolstered by qualitative interpretation.

Generalizability

Generalizability will be quite limited because analyses are restricted to a single site.

Organization and Management

We estimate that this study should require about six months to perform--one month for data preparation and five months for analysis and reporting. Once again, a prime contractor could subcontract with Georgia State for the data preparation. An estimate of required resources for a Georgia State study are given in Table 23.

5.7 School District of Philadelphia Data Base

5.7.1 Educational Models and Later Effects Hypotheses

Seven different FT models spanning a wide range of educational philosophies were implemented in Philadelphia. The model sponsors are: Bank Street, Kansas, SEDL, EDC, North Carolina, Harrison School, and Philadelphia (self-sponsored). Harrison School sponsors the Parent Implementation Model with program content identical to that of the Philadelphia Process Model sponsored by Philadelphia.

Table 23

Estimated Resources Required for the Proposed Georgia State Study

Prime Contract

Direct Costs

Computer	2,000
Other (telephone, Xerox, printing, graphics, consultants, travel)	3,000
Total	$5,000

Labor

Senior Analyst	6 months
Junior Analyst	3 months
Support Staff	3 months
Total	12 months

Subcontract

Data Preparation (Georgia State)

Total	$8,000

A wide variety of hypotheses can be made for these models, ranging from sleeper effects (Bank Street) to fade-out (Kansas). Studies already performed by the Philadelphia School District include quasi-longitudinal test performance studies extending to grade 7 and mobility studies. These studies show positive effects of low absence, Head Start, and some of the FT models, notably Kansas, Philadelphia Process, and Bank Street. Analyses were conducted through grade 7, where, as for grades K-3, the Kansas model appeared superior to others in securing academic performance gains on the California Achievement Test.

It should be noted that, with the possible exception of SEDL, none of the sponsors implemented in Philadelphia is recommended for study via other data sets. While this is beneficial in terms of enlarging the pool of sponsors on which some LE information is available, it is problematic from another viewpoint. Overlap among sponsors across data sets would have been particularly appealing in terms of dealing with issues of implementation, replicability, and transportability.

5.7.2 General Description of the Data Base

The Philadelphia Public Schools (Research and Evaluation Office) maintains an extensive, fully longitudinal file for all FT children. National evaluation Cohorts I-EK, II-EK, and III-EK are represented by about 2,100 child records per cohort, including only those children present in FT for at least five months during each of the four FT years. Children are traced through grade 8. Measures include full identification of educational programs in which children were enrolled (beginning in pre-K), attendance data, and the test score data listed in Table 24. The file includes children's names, phonetic names, race, sex, and date of birth. All seven FT models implemented in Philadelphia are represented on the data base. From 1974-75 on, data are available in computer files, but (because of accidental tape scratching) data from 1971-72, 1972-73, 1973-74 are available only in hard copy. In order to fully utilize the Philadelphia data, it would be necessary to regenerate these files by punching the records from hard copy or by merging data from the national evaluation.

The principal strengths of the data include easy access for computer-based studies, large sample size, longitudinal structure, and the presence of extensive educational treatment data (records of the programs in which FT and NFT children have participated) through grade 3. Weaknesses

Table 24

Test Data on Merged National Evaluation/Philadelphia File

MEASURE	COHORT	I									II						III							
	STREAM	EK									EK						EK							
	GRADE	K	1		2		3		4	5-8	K	1	2		3	4-8	K		1		2	3	3	4-7
	TEST POINT	F69 S70	F70	S71	F71	S S72	F72	S73	S74	W75-W78	F70 S71	S72	F72	S73	S74	W75-W79	F71	S72	F72	S73	S74	S75	W75	W76-W79
WRAT		X	X	X							X						X	X						
PSI		X									X						X							
LEE-CLARK		X									X						X							
ECI		X									X													
FACES		X	X	X							X						X							
SOCIAL INTEGRATION ITEMS		X	X	X							X													
MRT			X	X																				
SPONSOR ITEMS (VERBAL & QUANT.)			X	X																				
SPONSOR ITEMS (MISC.)			X	X																				
MAT PRIMER																		X	X					
MAT PRIMARY I												X	X							X				
MAT PRIMARY II					X	X	X							X							X			
MAT ELEMENTARY								X							X							X		
RAVEN					X			X							X							X		
COOPERSMITH					X			X							X							X		
IARS					X			X							X							X		
GUMPGOOKIES												X						X						
LOCUS OF CONTROL												X						X						
PPVT																	X	X						
CAT									X	X					X	X					X		X	X
SESAT																		X						

P6-6 B/13

Table 25

Process Data by Grade Available on the Philadelphia Data Base

Grade	Measure
Pre-K	Head Start Exposure
	Other Preschool Exposure
	Absence
	Special Ed Flag
K-3	School
	Grade
	Summer School Flag
	Absence
	FT Exposure
	Other Program Exposure
4-8	Absence
	School
	Grade

Table 26

Background Data by Cohort, Stream and Grade for the National Evaluation Data Base--To Be Merged with the Philadelphia Data Set*

	Cohort	I		II			III		IV
	Stream	EK	EF	EK		EF	EK	EF	EK
Variable	Grade	K,1,2,3	1,2,3	K,1,2	3	1,2,3	K,1,2,3	1,2,3	K,1,2
Project		x	x	x	x	x	x	x	x
School		x	x	x	x	x	x	x	x
FT/NFT designator		x	x	x	x	x	x	x	x
Grade		x	x	x	x	x			
Child birthdate		x	x	x	x	x	x	x	x
Sex		x	x	x	x	x	x	x	x
Ethnicity		x	x	x	x	x	x	x	x
First language spoken in the home		x	x	x	x	x	x	x	x
Second language spoken in the home		x	x	x	x	x	x	x	
Head Start flag		x	x	x	x	x			
Head Start equivalent flag					x				
Months Head Start or equivalent experience		x	x	x	x	x			
Months Head Start experience							x	x	
Months equivalent experience							x	x	
Days absent during year		x	x	x	x	x	x	x	x
Child eligible for FT services					x		x	x	x
Classroom service available					x		x	x	x
Lunch service available					x		x	x	x
Medical, dental service available					x		x	x	x
Other service available					x		x	x	x
Date child entered class					x				
Date child left class					x				
Years of data available on child							x	x	x
Parent interview history							x	x	x
Child test history							x	x	
Child FT/NFT by year							x	x	x
Months in FT prior to this year							x	x	x
Child entering grade									x

*Adapted from Middleton and Durgin (1978).

include lack of measures from K-1 (Cohort I) and K (Cohort II); lack of pre-tests (Fall administration), particularly for entering-K; lack of socioeconomic data; and lack of non-cognitive measures. These weaknesses can, to a large extent, be overcome by merging with the national evaluation data base.

The possibility of such a file merge was explored by Dr. Thomas McNamara of the Philadelphia Public School System, who determined in correspondence with SRI that merging national evaluation and Philadelphia files was feasible. In June 1977, SRI estimated (very roughly) a cost of $8,500 to merge third grade Metropolitan Achievement Test, Ravens, Intellectual Achievement Responsibility Scale, and Coopersmith scores for Cohorts I, II and III-EK with the Philadelphia files. A more extensive merge would, of course, cost proportionately more. We recommend a merge of grades K-3 for Cohorts I, II, III, and Cohort IV (which contains only Philadelphia and New York City children). The remainder of this preliminary study plan assumes that this merge will be made. The analyses would concentrate on Cohorts I-III. Note that all Philadelphia FT children are entering-K children.

Measures Available

Tables 24, 25, and 26 list the principal measures available in a merged national evaluation/Philadelphia file. School readiness was assessed by the Wide Range Achievement Test, the Peabody Picture Vocabulary Test, and the Preschool Inventory. Achievement in grades K-3 is measured by the Wide Range Achievement Test and the California Achievement Test, and in grades 4-8 by the California Achievement Test. In order to perform common metric longitudinal analyses, one would need to reduce these scores to a common metric, either by percentile matching or by a latent trait technique (e.g., use of the Rasch model).

Comparison Groups

The history of comparison groups in Philadelphia is complicated. An initial set of sixteen comparison schools was selected in 1969-70, but FT personnel felt that the income and education levels of these schools were poorly matched with the FT sample. Therefore, a new set of 44 schools was selected in 1970-71. By 1974-75, these schools had, one by one, been absorbed into other compensatory programs modeled after FT. The most extensive of these competing programs was a court-mandated expansion of FT-style curricula into NFT schools. This period of change in comparison schools (1969-75) corresponds to the span of FT years for Cohorts I through III. In spite of this, there should be an adequate subset of comparison groups that are not badly compromised for analytic purposes. However, analyses for later cohorts do not seem advisable.

It may be desirable to identify a second comparison group of children who are known to have received some alternative NFT treatment. Of course, to the extent that the competing treatments received by FT comparison children were more effective than regular schooling, the observed effects of FT will be attenuated. There are three reasons that such attenuation may be considered appropriate. First, most FT children are eligible for and would receive Title I services in the absence of FT. This is a reality in urban school systems and argues that children receiving Title I (not regular schooling) form the most appropriate comparison group for FT. Second, FT is more expensive than most other compensatory programs. It is therefore worthwhile to assess the relative effects of FT vs. other programs. Third, FT is in part designed to test and compare the effects of different models of education. There is good reason to consider Title I as another model of education (realizing that it differs from FT models in many important ways) and therefore include it in any evaluation as a valuable additional treatment.

Part of the richness of the Philadelphia data base is that it contains information on the competing educational treatments of NFT children on a year-to-year basis. As remarked above, it may be desirable to design a study using two NFT groups, one composed of children receiving competing treatments said to be exceptional in some way, and another composed of children receiving schooling not so designated. This would allow empirical estimates of the effectiveness of FT models as measured against two comparison groups, each representing a form of the reality of urban education.

We have not investigated in detail the comparability of these three groups (FT, other compensatory treatment, regular schooling) in terms of income and education, although for the first two groups the match is apparently reasonable. Although use of ANCOVA techniques would weaken conclusions in the eyes of many, they will probably be required. We recommend on the basis of our current knowledge that demographic matching of FT and comparison groups be assessed at both the sponsor and site level. The adequacy and meaning of the comparisons to be made should be explicitly laid out, and a statistical adjustment strategy should be selected.

Cross-sponsor comparisons would compare post-FT outcomes across sponsors. A variety of techniques could be used, ranging from simple statistical comparison of means to ANCOVA or other statistical adjustment techniques. Such analyses have been criticized and even ridiculed in the case of the national evaluation as "horse races," although other researchers (e.g., Bereiter and Kurland 1978) support such analyses. There is little question that such analyses would be unpopular with some of the sponsors.* Almost any presentation of sponsor-by-sponsor analyses would allow the readers to draw their own conclusions. Partly in order to avoid possibly mistaken reader interpretations, we recommend that comparisons be explicitly presented in a way that takes into account potential valid objections. The Philadelphia School District has taken the sponsor-by-sponsor approach in several of its own studies.

Demographic Data

Demographic data available are listed in Table 26.

Sample Sizes

The Philadelphia Continuance and Transfer Study (Report #7596) indicates that about 80% of the FT children in each of grades K-2 continue in FT the next year. A majority of the attrited 20% withdraw from the FT program, but not from the Philadelphia Public School System. Data obtained from Philadelphia indicate attrition of only about 3% per year from grade 3 to grade 8 and about 10% from grade 8 to grade 9. About 700 children per cohort are tracked through grade 6 and have city-wide test scores for grades 3-6 and were exposed to at least 3.5 years of FT. Any less stringent requirements on the sample would result in even higher sample sizes. Since the Philadelphia files are exceptionally clean, nearly all of these children are likely to be matchable with the national evaluation data base.** Thus, we expect about 700 longitudinal children per cohort distributed across models (total of 2,100) in the merged K-6 file. Extending the file to grade

---

*This is not to suggest that USOE should be guided by these objections.

**Note that as part of the national evaluation SRI tested nearly all children in the Philadelphia FT programs except those in the small parent-implemented program, whose classroom curriculum was the same as that in the Philadelphia Process Model (self-sponsored).

8 would result in further attrition of only about 6%. Even within individual models, sample sizes will be ample for most statistical purposes.

Other Technical Considerations

It should be acknowledged that several sponsors found that levels of model implementation were particularly low in Philadelphia. The large scale of the Philadelphia system (its operating budget is about $750 million annually), its vast management system, the importance of detailed teacher contracts, and militant teachers' unions all make implementation difficult. The situation is typical of large eastern school systems, however, and those represent a large fraction of the clients for whom FT was designed. While many analysts would prefer to separate implementation issues from program evaluations and attack these problems separately, this is not possible in a Philadelphia study. Analyses would have to consider levels of implementation of different models, especially if any cross-sponsor comparisons are made.

More recent cohorts of FT children present another problem. Since programs were funded in part by Title I, the federal government began to enforce Title I eligibility requirements for FT. Under these requirements, children who achieved at national median levels were forced to leave FT; i.e., success was penalized. Fortunately, this does not affect the Cohort I-III study sample.

5.7.3 Preliminary Study Plan

Preliminary Analyses

Several questions need to be answered to complete the design of a summative study plan.

- What are the primary sources of variance in post-FT performance measures? What is the relative importance of classroom, school, Head Start, and FT effects? What is implied for the choice of unit of analysis?

An important conclusion of the national evaluation was that "site effects" within model accounted for a large part of the total variance in outcomes. These site effects may include differences in effectiveness of project management, differences in level of implementation, differences among sites--i.e., geography, overall social and economic conditions, etc., differences in the match of educational philosophies between the sponsor and site, differences in the rate of personnel turnover, differences in educational treatment of the comparison group, and diffusion effects. Although Philadelphia is regarded as one "site," school and neighborhood effects involve similar sources of variance. The approximate contribution of school effects can be determined via a nested variance components analysis (Graybill 1961). The study would be limited by the small number of schools per sponsor but would still be possible. It would illuminate issues in model implementation, especially if it could be supported by anecdotal and other descriptive data.

If the variance component associated with school within sponsor is substantial, child-level hypothesis tests are not strictly valid because of the existence of serial correlations. These are the statistical manifestations of class-level effects on the children, who are not independent of each other within each class. Although most analyses must be conducted at the child level because of our interest in child-to-child differences and because of small samples for schools and classes within sponsor, it might be desirable to perform a few hypothesis tests at school level. Tests can also be constructed via error components regression models (Nerlove 1971). This subject was treated in the discussion of the Fordham data and by Singer and Goodrich (1979).

- A vast number of potential measures of performance are available. To keep the analytic effort within manageable scope, some selections must be made. How should these be made? What measures should be analyzed?

Selections of measures should be made according to sample size, reliability, and validity considerations. First, a routine psychometric analysis (alphas, inter-correlations, factor structure, etc.) of available measures should be conducted to help in selection of measures. A great deal has been written about validity of the measures used in USOE's national evaluation, and these opinions should be considered (e.g., McLean 1978). We would recommend that analyses concentrate on one or two school readiness measures. The post-FT achievement measure must be the California Achievement Test. The grades 1-3 measure must be constructed from a complex pattern of the Wide Range Achievement Test, the California Achievement Test, and numerous versions of the Metropolitan Achievement Test. Test schedules differ from cohort to cohort. To make comparisons simpler, we recommend that standardized scores be used, based on the FT/NFT population in Philadelphia. This could be criticized on the basis of differing test contents (the content of the Wide Range Achievement Test has many detractors), but would make analyses and organization of the study much simpler. The use of the Wide Range Achievement Test should be minimized except as a pretest and in grade 1, Cohort I (where no alternative is available). Objections by experts to the use of the Wide Range Achievement Test center on its use in later grades rather than grades K-1.

- What should be the role of the Ravens, Coopersmith, and the Intellectual Achievement Responsibility Scale? Are they better viewed as covariables not greatly subject to educational effects or as outcomes of the educational process?

The causal ambiguity of these three tests is troublesome and was part of the basis for criticism of the national evaluation by House et al. (1978). For LE studies, these measures (administered in spring of grade 3) would be used as independent variables, but one would still like more confidence in their interpretability. This could be accomplished by concentrating on Cohort I, the only case where these tests were given twice longitudinally (Fall 1971 and Spring 1973). Exiting Ravens, Coopersmith, and Intellectual Achievement Responsibility Scale scores would be regressed against Fall 1971 scores and demographic characteristics, and the adjusted means would be compared across sponsors and FT/NFT as measures of educational input. If little sponsor-to-sponsor variation in adjusted means exists, we would be prone to regard these three measures as covariables, i.e., as measures of the child's initial status. On the other hand, if there are interpretable sponsor effects, then they should be regarded as important independent measures in LE analyses. In any case, the role of these variables must be carefully chosen, taking into account the criticisms that have been made as part of the FT controversy.

- What should be the role of absence data in the model? Is absence more a measure of pre-existing conditions (e.g., family background), or is it an outcome of the educational program? Does absence cause poor academic performance, or is the reverse true? Does the use of socioeconomic status data add to the model?

Absence data have been examined in several sponsor studies, always with the same results--absence is negatively correlated with academic performance. But this does not inform us about its causal role. The Philadelphia data allow some limited tests to be made of the causal roles of variables for which time series are available. In this case, demographic data, absence, and standardized scores from achievement tests are available from grades K-8 across a large sample of children. A similar test is possible for the Fordham study and is discussed there. Since absence plays such a key role in several analyses, the results will have some general importance--if they are definitive.

Absence and retention data are important dependent variables in later grades, and they should be analyzed as life chances measures. It is their role as K-3 measures that appears puzzling.

- How can the comparison schools be classified in terms of demographic comparability and of competing NFT treatment? Is it feasible and desirable to select two or more comparison groups, both well matched on demographic characteristics but differing on type of NFT treatment received?

This point is discussed earlier, where the comparison groups are described.

- What should be the approach to the non-equivalent control group problem? What control variables should be selected? Should comparison groups be selected on a sponsor-by-sponsor basis, or should one large group be selected?

Comparison groups will probably be nearly equivalent to treatment groups for some sponsors. Since sponsors tend to be identified with subdistricts of Philadelphia, this can vary by neighborhood. The ideal would provide separate within-neighborhood comparison groups for each sponsor, but the erosion of comparison groups over the period of interest may make this unattainable technically. Answering the question, therefore, requires a detailed study.

ANCOVA should be used in any case. Even when comparison groups are closely matched, ANCOVA is useful in achieving greater statistical power and, often, reduced heterogeneity of variance. Selection of covariables for ANCOVA is a routine matter. The option of different covariable models for different sponsors may be required if there appears to be substantial heterogeneity of regression on demographic variables.

Summative Analyses

The domain of demographic and K-3 measures, which are independant variables for LE analyses, should be explored by the same basic methods discussed for Oregon. However, analyses would be conducted within sponsor. Results should be presented by sponsor as well, where comparisons may be with local, sponsor-specific control groups, selected to represent neighborhood conditions, as well as with a pooled NFT group. As discussed earlier, results would indeed be presented in sponsor-by-sponsor format, since these inter-sponsor comparisons appear to be inevitable.

Specific questions are as follows.

- What is the relation of the entering-K data panel to initial status (service eligibility, ethnicity, sex, age) and to HS experience? What is the relation of the spring-K panel to the entering-K panel, initial status, Head Start experience, and FT experience?

Data panels are different from cohort to cohort, so analyses must be separate. This step may help in excluding irrelevant measures and may thus simplify the analyses. The objective is to obtain causal multivariate connections that explain the measures. It is the first step in building a cumulative panel data model.

FT experience can be differentiated by model. As the analysis proceeds grade by grade, the development of model differences through time should be examined. There may have to be entirely separate models by sponsor, although the causal interrelation of measures shouldn't differ substantially across models.

- What is the relation of the fall-grade 1 panel to foregoing panels? Of spring-grade 1 to fall-grade 1 and foregoing panels? What earlier data drop out (replaced by later information)?
- What are the similar models that connect grades 2 and 3 to prior data?

These three sets of analyses will yield a clean structural model that shows the causal evolution of the data panel through the FT years. Many variables will probably be dropped from the model on grounds of statistical irrelevance. The overall grades K-3 model, whose structure is determined by ordinary regression as discussed, should then be re-estimated via LISREL. Substantive theory should be used to support model building throughout the process.

For the following questions, only the California Achievement Test is available as an outcome measure. The most potent variables from the grades K-3 panels will be taken as independent variables in a new set of regressions.

This step is descriptive in nature. Many techniques can be tried, including time-series stochastic models, curve fitting (usually involving the logistic curve), principal-component techniques, etc. Most of these techniques come from the research on human growth.

- How does growth in grades 4-8 on California Achievement Test scores (as parameterized) depend upon initial status measures, important independent variables from grades K-3, Head Start and FT experience (including duration), absence, retention, mobility, etc.?
- How do retention and absences in grades 4-8 relate to the same independent variables?

These questions would be answered via standard multiple regression techniques. Of course, the dependence on specific FT models would be examined.

Validity

Traditional analyses, especially least squares techniques like multiple regression, provide little protection against distortion of the analyses by unusual behavior in subsets of the data. In the Philadelphia study, we have a great number of subjects over three cohorts, seven models, and several years. Cohort-to-cohort differences, year-to-year differences, etc., must be searched for and intepreted. If necessary, data must be discarded. A particularly useful technique in performing these analyses is the jackknife (Mosteller and Tukey 1977). We would recommend consideration of this and other robust methods if resources permit.

Generalizability

The universe of generalizability is the large eastern city school system. Philadelphia shares with other such sites characteristics such as large ethnic populations, organization and management on an enormous scale, many special interest groups, militant teachers' unions, and considerable inertia for organizational change. The Philadelphia analysis should take into account anecdotal and quantitative evidence that speaks to the problem of generalizing results to other cities.

Organization and Management

USOE has indicated that they favor letting a separate contract for preparation of a Philadelphia data file.* Assuming such work is accomplished, funding of a Philadelphia study contract would involve no substantial further data preparation. From the time a file is ready, the study could be performed in a period of about one year. Resource estimates (excluding data preparation) are given in Table 27.

---

*Such a contract has been let to Philadelphia and SRI, and the merge is proceeding.

5.8 Integration of Studies

We recommend that, at the conclusion of whichever sponsor-level studies are performed, the results be summarized and integrated with each other and with existing sponsor and site reports of post-FT effects. Attempts should be made to generalize findings and to adopt plausible hypotheses about later effects of FT. Clearly, a good deal of subjective judgment must enter into such a project, but we think the results will be useful for organizing knowledge about FT and as a means to determine new directions of research.

Table 27

Estimated Resources Required for the Proposed Philadelphia Study

Direct Costs	
Computer	$100,000
Other (telephone, Xerox printing, graphics, consultants, travel)	20,000
Total	$120,000
Labor	
Senior Analysts	24 months
Junior Analysts	24 months
Support Staff	24 months
Total	72 months

133

References

Abelson, W.C. "Head Start Graduates in School: Studies in New Haven, Connecticut." In S. Ryan (Ed.), A Report on Longitudinal Evaluations of Preschool Programs. Vol. 1, Longitudinal Evaluations. Washington, D.C.: Depart of Health, Education and Welfare, Publication No. (OHD) 74-24, 1974.

Becker, W.C. and S. Engelmann. "Analysis of Achievement Data on Six Cohorts of Low Income Children from 20 School Districts in the University of Oregon Direct Instruction Follow Through Model." Technical Report 78-1, University of Oregon Follow Through Project, Eugene, Oregon, December 1978.

Belsley, D.A., E. Kuh, and R.E. Welsch. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley & Sons, forthcoming.

Bereiter, C. and M. Kurland. "A Constructive Look at Follow Through Results." Unpublished manuscript. Toronto, Canada: Ontario Institute for Studies in Education, October 1978

Boardman, A.E. and R.J. Murnane. "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." Sociology of Education, 52, 113-121, April 1979.

Bridewell, J.L. and R.H. Edwards. "Follow Through 1977-1978," paper presented at 1979 meeting of the American Educational Research Association, San Francisco, April 1979.

Caines, P.E. "Weak and Strong Feedback--Free Processes." IEEE Trans. Automat. Contr., AC-21, 5, October 1976.

Bryk, A.S. and H.I. Weisberg. "Use of the Nonequivalent Control Group Design When Subjects are Growing." Psychological Bulletin, Vol. 84, No. 5, 950-962, 1977.

Carlson, K., unpublished work in progress, Cambridge, MA: Abt Associates Inc., 1979.

Chow, G.C. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," Econometrica, 28, July 1960.

Cloud, K., R.K. Rentfrow and J. Hildebrandt. "Later School Success of Follow Through Graduates in Lakewood, New Jersey Public Schools." Tucson, AZ: Center for Educational Research and Development, University of Arizona, July 1979.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press, 1977.

Cronbach, L.J., D.R. Rogosa, R.E. Floden and G.G. Price. "Analysis of Covariance in Non-Randomized Experiments: Parameters Affecting Bias." Stanford Evaluation Consortium, August 1977.

References

Draper, N.R. and Smith, H. Applied Regression Analysis. New York: Wiley & Sons, 1966.

Duncan, O.D. Introduction to Structural Equation Models. New York: Academic Press, 1975.

Dy Reyes, F. "A Test of the Direction of Causality Between Money and Income in Canada, Japan and the United States." Ph.D. Dissertation Dept. Econ., Iowa State University, Ames, 1974.

Edwards, R.H. and J.L. Bridewell. "Following Through on Follow Through." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Feige, E. and D.K. Pearce. "The Causality Relationship Between Money and Income: A Time Series Approach." Paper presented at the Midwest Economic Conference, Chicago, IL, 1974.

Garet, M. Follow Through Evaluation Studies in the Next Two Years: Ten Potential Plans. Cambridge, MA: Huron Institute, 1977.

Goldberger, A.S. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations. "University of Wisconsin, Institute for Research on Poverty: Discussion Paper 123-72, April 1972.

Goodrich, R.L. and J. Singer, "Cognitive Change in the NDCS," Unpublished manuscript. Cambridge, MA: Abt Associates Inc., 1978.

Goodrich, R.L. and R.G. St.Pierre. The SRI Data Base: Opportunities for Studying Later Effects of Follow Through. AAI Report #79-1, to U.S. Office of Education pursuant to Contract No. 300-78-0443. Cambridge, MA: Abt Associates Inc., March 1979.

Goodrich, R.L. and R.G. St.Pierre. Sponsor and Site Data Bases: Opportunities for Studying Later Effects of Follow Through. AAI Report #79-75, to U.S. Office of Education pursuant to Contract No. 300-78-0443. Cambridge, MA: Abt Associates Inc., June 1979.

Granger, C.W.J. "Economic Processes Involving Feedback." Inform. Contr., 6, p. 28-48, 1963.

Granger, C.W.J. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." Econometrica, 37, July 1969.

Graybill, F.A. An Introduction to Linear Statistical Models. New York; McGraw-Hill, 1961.

Greenhouse, S.W. and Geisser, S. "On Methods in the Analysis of Profile Data," Psychometrika, 24, 2, 1959.

References

Haney, W. The Follow Through Planned Variation Experiment, Vol. V: A Technical History of the Nation Follow Through Evaluation. Cambridge, MA: The Huron Institute, August 1977.

Hartman, S.A. and H.D. Acland. "Final Report on the Follow Through City Data." Cambridge, MA: The Huron Institute, June 1973.

Hoephner, Zagorski, R.H. and J. Wellisch. The Sample for the Sustaining Effects Study and Projections of its Characteristics to the National Population. Report to the U.S. Office of Education pursuant to Contract No. OE-300-75-0332. Santa Monica, CA: System Development Corporation, June 1977.

House, E., G. Glass, D. Walker, and L. McLean. "No Simple Answer: Critique of a Follow Through Evaluation." Harvard Educational Review, 1978, 48, (2), 128-160.

Joreskog, K.G. "An Economic Model for Multivariate Panel Data." Annales de l'Insee, 30-31, April-September 1978.

Joreskog, K.G. "Structural Equation Models in the Social Sciences: Specification, Estimation and Testing." In P.R. Krishnaiah, Ed., Applications of Statistics. Amsterdam: North Holland Publishing Co., 265-287, 1977.

Lazar, I., V.R. Hubbell, H. Murray, M. Rosche and J. Royce. Summary: The Persistance of Preschool Effects. Report to the Administration for Children Youth and Families pursuant to Grant No. 18-76-07843. Ithaca, NY: Cornell University, 1977.

Lord, F.M. "Large-Scale Covariance Analysis When the Control Variable is Fallible." Journal of the American Statistical Association, 1960, 55, 307-520.

Madaus, G.F. and T. Marx. "Report on the Follow Through Later Grades Study." Cambridge, MA: Huron Institute, April 1973.

Magidson, J. "Toward a Causal Model Approach for Adjusting for Preexisting Differences in the Nonequivalent Congrol Group Situation: A General Alternative to ANCOVA." Evaluation Quarterly, Vol. 1, No. 3, August 1977.

McLean, L. "Evaluation of Early Childhood Education--No Simple Answers to the Right Questions." Paper presented at the conference "Nature, Nurturance and School Achievement," York University, Toronto, Canada. May 8-10, 1978.

Middleton, T. and L.A. Durgin. Summary of Follow Through Data. Report to the U.S. Office of Education pursuant to Contracts Nos. OEC-0-522480-4633 (100) and OEC-0-0582. Menlo Park, CA: SRI International, April 1978.

References

Mosteller, F. and J.W. Tukey. Data Analysis and Regression. Reading, MA: Addison-Wesley, 1977.

Nerlove, M. "A Note on Error Components Models." Econometrica, 39, 2, p. 383, 1971.

Porter, A.C. "The Effects of Using Fallible Variables in the Analysis of Covariance." Unpublished doctoral dissertation, University of Wisconsin, 1967.

Ramp, E., R. Williams, and A. Branden. Planning Information for Follow Through Experiments. Proposal submitted to the U.S. Office of Education in response to RFP 78-101. Lawrence, KS: University of Kansas, 1978.

Riley, G.R. "Follow Through Program Evaluation, 1976-77." Wichita Public Schools Division of Research, Planning and Development Services, January 1978.

Seitz, V., N.H. Apfel and C. Efron. "Long-term Effects of Early Intervention: The New Haven Project." Paper presented at the annual meeting of the American Association for the Advancement of Science, Denver, February 1977.

Sims, G.A. "Money, Income and Causality," American Economic Review, 62, 1972, p. 540-552.

Singer, J. and R.L. Goodrich. "Aggregation and the Unit of Analysis Issue in the National Day Care Study." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Stebbins, L.B., R.G. St.Pierre, E.C. Proper, R.B. Anderson, T.R. Cerva. Education as Experimentation: A Planned Variation Model. Vol. IV-A of the Follow Through Final Report. Cambridge, MA: Abt Associates Inc., April 1977.

Tukey, J.W. Exploratory Data Analysis. Reading, MA: Addison-Wesley, 1977.

Weber, B. and M. Fuhrmann. "A Study of District A's Former Follow Through Students' Retention of Basic Skills After Six Years Out of the Program," April 1978. Paper prepared for "District A."

U.S. Office of Education. "An Examination of Alternatives for Follow Through Experimentation." RFP 78-106. Washington, D.C.: 1978.

Wall, K.D. "An Application of Simultaneous Estimation to the Determination of Causality Between Money and Income." Programme for Research into Econometric Methods, University of London, Discussion Paper 8. London, England, 1974.

Wolff, M. and A. Stein. Six Months Later, Head Start Evaluation Project. New York: Yeshiva University, Ferkauf Graduate School of Education, 1966.

References

Zajonc, R.B. "Birth Order and Intelligence: Dumber by the Dozen." Psychology Today, 1975, 8(8), 37-44.

Zajonc, R.B. "Family Configuration and Intelligence." Science, 1976, 192, 227-236.

Appendix 1

The Linear Discriminant Function as a Composite Covariate

Assume that a univariate outcome variable y is explained by a "true" model of the form

$$y = aT + x + e, \qquad (1)$$

where y, T, x, e are n-by-1 matrices (vectors), and n is the sample size. T, x, and e represent the treatment (dichotomous), the "true" covariate, and error (uncorrelated with T or x), respectively. Assume that the nature of the true covariate is unknown, but that there is available a stock of p covariates that form the components of the n-by-p matrix Z. The point is to estimate the value of the treatment effect a even though x is unknown.

Let Z represent the space generated by the stock of covariables. Then x can be represented as the sum of its projection on Z and a component u orthogonal to Z, i.e.,

$$x = x^* + u,$$

where

$$u \perp Z,$$

and specifically,

$$u \perp x^*.$$

In case x actually lies in Z, then u = 0 and we call Z complete. Make the (restrictive) assumption that

$$T \perp u,$$

i.e., the covariance of T and u is zero. This is clearly true when Z is complete and is reasonable in most other cases. Since x is unknown, this assumption cannot be checked, but by using as many relevant covariables as possible the magnitude of u and hence the magnitude of cov(T,u) is minimized.

One method of estimating the treatment effect a in Equation (1) is to use the regression model:

$$y = aT + zB + e, \qquad (2)$$

estimating a and B by ordinary least squares. It is not difficult to show that under our restrictive assumption the ordinary least-squares estimate $\hat{a}$ is

Appendix 1

(continued)

consistent, even when Z is incomplete, i.e., the space of available regressors need not include the true covariate to yield a consistent treatment-effect estimate.

The problem is that a large number of covariables may be required in order to obtain an unbiased estimate of a. This leads to two problems:

(1) When the total N is small (as usually happens when cases are aggregated), the degrees of freedom absorbed in estimating B seriously reduce the statistical power of hypothesis tests.

(2) T and Z are usually strongly multi-collinear, so the treatment-effect estimate has a large error variance. Estimates are likely to be heterogeneous across sites or other data subsets.

An alternative approach is to precompute a composite covariable $W = ZC$ (where C is a p-by-1 matrix) and to use it as a univariate covariable in the model

$$y = aT + bW + e. \qquad (3)$$

In the unit-weighted composite covariate (UWCC) method the weighting vector C consists of the p-by-1 vector of reciprocal standard deviations of the components of Z. Proponents of the technique assert* that the method is more robust and not seriously less powerful than ordinary least squares. The method suggested here has these advantages plus one more--the estimate for a in Equation (3) is unbiased, even though the "true" model remains unknowable.

Select w as the projection of T on Z, i.e.,

$$W = \Omega T, \qquad (4)$$

where the projection operator Ω is defined by

$$\Omega = Z(Z'Z)^{-1}Z'. \qquad (5)$$

When homogeneity assumptions are satisfied, W is proportional to the linear discriminant function but is in fact more general. Under the assumptions stated, the least-squares estimate $\hat{a}$ in Equation (3) is an unbiased estimator of a in Equation (1).

---

*Communicated to us by Jacob Cohen. He cites Robin Dawes and Bert Green as proponents of UWCC.

Appendix 1

(continued)

Proof of Unbiasedness

The ordinary least-squares estimate for a in the model given by Equation (3) is:

$$\hat{a} = (T'TW'W - (T'W)^2)^{-1} (W'WT'y - T'WW'y) \tag{6}$$

Using Equation (1), it is possible to obtain:

$E\hat{a} = a + \Delta$, where the bias Δ is given by

$$\Delta = (T'TW'W - (T'W)^2)^{-1} (W'WT'x - T'WW'x) \tag{7}$$

Now substitute Equation (4) into this formula to obtain:

$$\Delta = \left[T'(I-\Omega)T\right]^{-1} \left[T'(I-\Omega)x\right] . \tag{8}$$

Since $(I-\Omega)x = u$,

$$\Delta = (T'(I-\Omega)T)^{-1} T'u$$

and the estimate $\hat{a}$ is unbiased if and only if $T'u = 0$, which is true by assumption.

Conclusion

We have shown that unbiased treatment effect estimates can be obtained by using the (univariate) composite $Z(Z'Z)^{-1}Z'T = \Omega T$ as a covariable. It is not necessary to view this covariate as an estimate of any "true" variable. Thus, its estimation does not absorb degrees of freedom. The result is easy to extend to the case of several treatment groups with heterogeneous "true" models.